# Physics 2000

**E. R. Huggins** Dartmouth College





physics2000.com

## MKS Units (link to CGS Units)

m = meters	kg = kilog	grams	s = seconds			
N = newtons	J = joules		C = coulombs			
T = tesla	F = farads	S	H = henrys	Powers	of 10	
A = amperes	K = kelvi	ns	mol = mole		Dut	•
speed of light		c	$3.00 \times 10^8 \mathrm{m}$ / s	Power	Prefix	Symbol
gravitational con	nstant	G	$6.67 \times 10^{-11} \text{N} \cdot \text{m}^2 / \text{kg}^2$	$10^{12}$	tera	Т
permittivity con	stant	80	$8.85 \times 10^{-12} \text{F/m}$	10'	giga	G
	Juint	0		10 <sup>6</sup>	mega	Μ
permeability con	nstant	$\mu_0$	$1.26 \times 10^{-6} \text{H/m}$	$10^{3}$	kilo	k
elementary char	ge	e	$1.60 \times 10^{-19}$ C	$10^{2}$	hecto	h
electron volt		eV	$1.60 \times 10^{-19} \text{J}$	$10^{-1}$	deci	d
electron rest ma	SS	m <sub>e</sub>	$9.11 \times 10^{-31} \text{kg}$	$10^{-2}$	centi	с
proton rest mass	5	m <sub>p</sub>	$1.67 \times 10^{-27} \text{kg}$	$10^{-3}$	milli	m
Planck constant		h	$6.63 \times 10^{-34} \text{J} \cdot \text{s}$	10 <sup>-6</sup>	micro	μ
Planck constant	/2π	ħ	$1.06 \times 10^{-34} \text{J} \cdot \text{s}$	10 <sup>-9</sup>	nano	n
Bohr radius		r <sub>b</sub>	$5.29 \times 10^{-11} m$	$10^{-12}$	pico	р
Bohr magneton		$\mu_{b}$	$9.27 \times 10^{-24}$ J / T	$10^{-15}$	femto	f
Boltzmann cons	stant	k	$1.38 \times 10^{-23}$ J/K			
Avogadro const	ant	N <sub>A</sub>	$6.02 \times 10^{23} \text{mol}^{-1}$			
universal gas co	onstant	R	8.31 J/mol·K			

#### Dimensions

Quantity	Unit		Equivalents	
Force	newton	Ν	J/m	kg•m/ $s^2$
Energy	joule	J	N•m	$kg \cdot m^2/s^2$
Power	watt	W	J/s	kg•m <sup>2</sup> /s <sup>3</sup>
Pressure	pascal	Ра	N/m <sup>2</sup>	$kg/m \cdot s^2$
Frequency	hertz	Hz	cycle/s	s <sup>-1</sup>
Electric charge	coulomb	С		A•s
Electric potential	volt	V	J/C	kg•m <sup>2</sup> /A•s <sup>3</sup>
Electric resistance	ohm	Ω	V/A	kg•m <sup>2</sup> /A <sup>2</sup> •s <sup>3</sup>
Capacitance	farad	F	C/V	$A^2 \cdot s^4 / kg \cdot m^2$
Magnetic field	tesla	Т	N•s/C•m	$kg/A \cdot s^2$
Magnetic flux	weber	Wb	$T \cdot m^2$	kg•m <sup>2</sup> /A•s <sup>2</sup>
Inductance	henry	Н	V•s/A	kg•m <sup>2</sup> /A <sup>2</sup> •s <sup>2</sup>

## Physics2000



Student project by Bob Piela explaining the hydrogen molecule ion.

by E. R. Huggins Department of Physics Dartmouth College Hanover, New Hampshire

#### Preface

#### **ABOUT THE COURSE**

Physics2000 is a calculus based, college level introductory physics course that is designed to include twentieth century physics throughout. This is made possible by introducing Einstein's special theory of relativity in the first chapter. This way, students start off with a modern picture of how space and time behave, and are prepared to approach topics such as mass and energy from a modern point of view.

The course, which was developed during 30 plus years working with premedical students, makes very gentle assumptions about the student's mathematical background. All the calculus needed for studying Physics2000 is contained in a supplementary chapter which is the first chapter of a *physics based* calculus text. We can cover all the necessary calculus in one reasonable length chapter because the concepts are introduced in the physics text and the calculus text only needs to handle the formalism. (The remaining chapters of the calculus text introduce the mathematical tools and concepts used in advanced introductory courses for physics and engineering majors. These chapters will appear on a later version of the *Physics2000* CD, hopefully next year.)

In the physics text, the concepts of velocity and acceleration are introduced through the use of strobe photographs in Chapter 3. How these definitions can be used to predict motion is discussed in Chapter 4 on calculus and Chapter 5 on the use of the computer.

Students themselves have made major contributions to the organization and content of the text. Student's enthusiasm for the use of Fourier analysis to study musical instruments led to the development of the MacScope<sup>™</sup> program. The program makes it easy to use Fourier analysis to study such topics as the normal modes of a coupled aircart system and how the energytime form of the uncertainty principle arises from the particle-wave nature of matter. Most students experience difficulty when they first encounter abstract concepts like vector fields and Gauss' law. To provide a familiar model for a vector field, we begin the section on electricity and magnetism with a chapter on fluid dynamics. It is easy to visualize the velocity field of a fluid, and Gauss' law is simply the statement that the fluid is incompressible. We then show that the electric field has mathematical properties similar to those of the velocity field.

The format of the standard calculus based introductory physics text is to put a chapter on special relativity following Maxwell's equations, and then put modern physics after that, usually in an extended edition. This format suggests that the mathematics required to understand special relativity may be even more difficult than the integral-differential equations encountered in Maxwell's theory. Such fears are enhanced by the strangeness of the concepts in special relativity, and are driven home by the fact that relativity appears at the end of the course where there is no time to comprehend it. This format is a disaster.

Special relativity does involve strange ideas, but the mathematics required is only the Pythagorean theorem. By placing relativity at the beginning of the course you let the students know that the mathematics is not difficult, and that there will be plenty of time to become familiar with the strange ideas. By the time students have gone through Maxwell's equations in Physics2000, they are thoroughly familiar with special relativity, and are well prepared to study the particle-wave nature of matter and the foundations of quantum mechanics. This material is not in an extended edition because there is of time to cover it in a comfortably paced course.

#### ABOUT THE PHYSICS2000 CD

The *Physics2000* CD contains the complete *Physics2000* text in Acrobat<sup>TM</sup> form along with a supplementary chapter covering all the calculus needed for the text. Included on the CD is a motion picture on the time dilation of the Muon lifetime, and short movie segments of various physics demonstrations. Also a short cookbook on several basic dishes of Caribbean cooking. The CD is available at the web site

#### www.physics2000.com

The cost is \$10.00 postpaid.

Also available is a black and white printed copy of the text, including the calculus chapter and the CD, at a cost of \$ 39 plus shipping.

The supplementary calculus chapter is the first chapter of a *physics based* calculus text which will appear on a later edition of the *Physics2000* CD. As the chapters are ready, they will be made available on the web site.

#### **Use of the Text Material**

Because we are trying to change the way physics is taught, Chapter 1 on special relativity, although copyrighted, may be used freely (except for the copyrighted photograph of Andromeda and frame of the muon film). All chapters may be printed and distributed to a class on a non profit basis.

#### **ABOUT THE AUTHOR**

E. R. Huggins has taught physics at Dartmouth College since 1961. He was an undergraduate at MIT and got his Ph.D. at Caltech. His Ph.D. thesis under Richard Feynman was on aspects of the quantum theory of gravity and the non uniqueness of energy momentum tensors. Since then most of his research has been on superfluid dynamics and the development of new teaching tools like the student built electron gun and MacScope<sup>TM</sup>. He wrote the non calculus introductory physics text Physics1 in 1968 and the computer based text Graphical Mechanics in 1973. The Physics2000 text, which summarizes over thirty years of experimenting with ways to teach physics, was written and class tested over the period from 1990 to 1998. All the work of producing the text was done by the author, and his wife, Anne Huggins. The text layout and design was done by the author's daughter Cleo Huggins who designed eWorld<sup>TM</sup> for Apple Computer and the Sonata<sup>TM</sup> music font for Adobe Systems.

The author's eMail address is

#### lish.huggins@dartmouth.edu

The author is glad to receive any comments.

### **Table of Contents**

#### PART 1

#### **Front Cover**

MKS Units	Front cover-2
Dimensions	Front cover-2
Powers of 10	Front cover-2

#### Preface

About the Course	. iii
About the Physics2000 CD	. iv
Use of the Text Material	. iv
About the Author	. iv

#### **INTRODUCTION—AN OVERVIEW OF PHYSICS**

Space And Time	int-2
The Expanding Universe	int-3
Structure of Matter	int-5
Atoms	int-5
Light	int-7
Photons	int-8
The Bohr Model	int-8
Particle-Wave Nature of Matter	int-10
Conservation of Energy	int-11
Anti-Matter	int-12
Particle Nature of Forces	int-13
Renormalization	int-14
Gravity	int-15
A Summary	int-16
The Nucleus	int-17
Stellar Evolution	int-19
The Weak Interaction	int-20
Leptons	int-21
Nuclear Structure	int-22
A Confusing Picture	int-22
Quarks	int-24
The Electroweak Theory	int-26
The Early Universe	int-27
The Thermal Photons	int-29

#### CHAPTER 1 PRINCIPLE OF RELATIVITY

The Principle of Relativity	1-2
A Thought Experiment	1-3
Statement of the Principle of Relativity	1-4
Basic Law of Physics	1-4
Wave Motion	1-6
Measurement of the Speed of Waves	1-7
Michaelson-Morley Experiment	1-11
Einstein's Principle of Relativity	1-12
The Special Theory of Relativity	1-13
Moving Clocks	1-13
Other Clocks	1-18
Real Clocks	1-20
Time Dilation	1-22
Space Travel	1-22
The Lorentz Contraction	1-24
Relativistic Calculations	1-28
Approximation Formulas	1-30
A Consistent Theory	1-32
Lack of Simultaneity	1-32
Causality	1-36
Appendix A	1-39
Class Handout	1-39

#### CHAPTER 2 VECTORS

Vectors	2-2
Displacement Vectors	2-2
Arithmetic of Vectors	2-3
Rules for Number Arithmetic	2-4
Rules for Vector Arithmetic	2-4
Multiplication of a Vector by a Number	2-5
Magnitude of a Vector	2-6
Vector Equations	2-6
Graphical Work	2-6
Components	2-8
Vector Equations in Component Form	2-10
Vector Multiplication	2-11
The Scalar or Dot Product	2-12
Interpretation of the Dot Product	2-14
Vector Cross Product	2-15
Magnitude of the Cross Product	2-17
Component Formula for the Cross Product	2-17
Right Handed Coordinate System	2-18

#### CHAPTER 3 DESCRIPTION OF MOTION

Displacement Vectors	. 3-5
A Coordinate System	. 3-7
Manipulation of Vectors	. 3-8
Measuring the Length of a Vector	. 3-9
Coordinate System and Coordinate Vectors	3-11
Analysis of Strobe Photographs	3-11
Velocity	3-11
Acceleration	3-13
Determining Acceleration	
from a Strobe Photograph	3-15
The Acceleration Vector	3-15
Projectile Motion	3-16
Uniform Circular Motion	3-17
Magnitude of the Acceleration for Circular Motion	3-18
An Intuitive Discussion of Acceleration	3-20
Acceleration Due to Gravity	3-21
Projectile Motion with Air Resistance	3-22
Instantaneous Velocity	3-24
Instantaneous Velocity from a Strobe Photograph	3-26

#### CHAPTER 4 CALCULUS IN PHYSICS

Limiting Process	4-1
The Uncertainty Principle	4-1
Calculus Definition of Velocity	4-3
Acceleration	4-5
Components	4-6
Distance, Velocity and	
Acceleration versus Time Graphs	4-7
The Constant Acceleration Formulas	4-9
Three Dimensions	4-11
Projectile Motion with Air Resistance	4-12
Differential Equations	4-14
Solving the Differential Equation	4-14
Solving Projectile Motion Problems	4-16
Checking Units	4-19

## CHAPTER 5 COMPUTER PREDICTION OF MOTION

Step-By-Step Calculations	5-1
Computer Calculations	5-2
Calculating and Plotting a Circle	5-2
Program for Calculation	5-4
The DO LOOP	5-4
The LET Statement	5-5
Variable Names	5-6
Multiplication	5-6
Plotting a Point	5-6
Comment Lines	5-7
Plotting Window	5-7
Practice	5-8
Selected Printing (MOD Command) 5	j-10
Prediction of Motion 5	-12
Time Step and Initial Conditions 5	-14
An English Program for Projectile Motion 5	-16
A BASIC Program for Projectile Motion 5	j-18
Projectile Motion with Air Resistance 5	j-22
Air Resistance Program 5	-24

#### CHAPTER 6 MASS

Definition of Mass	6-2
Recoil Experiments	6-2
Properties of Mass	6-3
Standard Mass	6-3
Addition of Mass	6-4
A Simpler Way to Measure Mass	6-4
Inertial and Gravitational Mass	6-5
Mass of a Moving Object	6-5
Relativistic Mass	6-6
Beta ( <b>β</b> ) Decay	6-6
Electron Mass in $oldsymbol{eta}$ Decay	6-7
Plutonium 246	6-8
Protactinium 236	6-9
The Einstein Mass Formula	6-10
Nature's Speed Limit	6-11
Zero Rest Mass Particles	6-11
Neutrinos	6-13
Solar Neutrinos	6-13
Neutrino Astronomy	6-14

## CHAPTER 7 CONSERVATION OF LINEAR & ANGULAR MOMENTUM

Conservation of Linear Momentum 7-2
Collision Experiments
Subatomic Collisions 7-7
Example 1 Rifle and Bullet 7-7
Example 2 7-8
Conservation of Angular Momentum 7-9
A More General Definition of Angular Momentum 7-12
Angular Momentum as a Vector
Formation of Planets 7-17

#### CHAPTER 8 NEWTONIAN MECHANICS

Force	8-2
The Role of Mass	8-3
Newton's Second Law	8-4
Newton's Law of Gravity	8-5
Big Objects	8-5
Galileo's Observation	8-6
The Cavendish Experiment	8-7
"Weighing" the Earth	8-8
Inertial and Gravitational Mass	8-8
Satellite Motion	8-8
Other Satellites	8-10
Weight	8-11
Earth Tides	8-12
Planetary Units	8-14
Table 1 Planetary Units	8-14
Computer Prediction of Satellite Orbits	8-16
New Calculational Loop	8-17
Unit Vectors	8-18
Calculational Loop for Satellite Motion	8-19
Summary	8-20
Projectile Motion Program	0-20 9 21
Orbit-1 Program	8_21
Satellite Motion Laboratory	8-23
Kenler's Laws	8-24
Kenler's First I aw	8-26
Kepler's Second Law	8-27
Kepler's Third Law	8-28
Modified Gravity and General Relativity	8-29
Conservation of Angular Momentum	8-32
Conservation of Energy	8-35
control valient of Energy	0 00

## CHAPTER 9 APPLICATIONS OF NEWTON'S SECOND LAW

Addition of Forces
Spring Forces
The Spring Pendulum
Computer Analysis of the Ball Spring Pendulum 9-8
The Inclined Plane 9-10
Friction
Inclined Plane with Friction
Coefficient of Friction 9-13
String Forces
The Atwood's Machine 9-16
The Conical Pendulum
Appendix: The ball spring Program 9-20

#### CHAPTER 10 ENERGY

· · · · · · · · · · · · · · · · · · ·	10-1
Conservation of Energy	10-2
Mass Energy	10-3
Ergs and Joules	10-4
Kinetic Energy	10-5
Example 1	10-5
Slowly Moving Particles	10-6
Gravitational Potential Energy	10-8
Example 2	. 10-10
Example 3	10-11
Work	. 10-12
The Dot Product	. 10-13
Work and Potential Energy	10-14
Non-Constant Forces	10-14
Potential Energy Stored in a Spring	10-16
Work Energy Theorem	. 10-18
Several Forces	. 10-19
Conservation of Energy	10-20
Conservative and Non-Conservative Forces	10-21
Gravitational Potential Energy on a Large Scale	10-22
Zero of Potential Energy	10-22
Gravitational PotentialEnergy in a Room	10-25
Satellite Motion and Total Energy	. 10-26
Example 4 Escape Velocity	10-28
Black Holes	. 10-29
A Practical System of Units	10-31

#### CHAPTER 11 SYSTEMS OF PARTICLES

Center of Mass 11-2
Center of Mass Formula 11-3
Dynamics of the Center of Mass 11-4
Newton's Third Law 11-6
Conservation of Linear Momentum 11-7
Momentum Version of Newton's Second Law 11-8
Collisions 11-9
Impulse 11-9
Calibration of the Force Detector 11-10
The Impulse Measurement 11-11
Change in Momentum 11-12
Momentum Conservation during Collisions 11-13
Collisions and Energy Loss 11-14
Collisions that Conserve Momentum and Energy 11-16
Elastic Collisions 11-17
Discovery of the Atomic Nucleus 11-19
Neutrinos 11-20
Neutrino Astronomy 11-21

#### CHAPTER 12 ROTATIONAL MOTION

Radian Measure Angular Velocity	12-2 12-2 12-3
Angular Analogy	. 12-3
Tangential Distance, Velocity and Acceleration .	12-4
Radial Acceleration	12-5
Bicycle Wheel	12-5
Angular Momentum	12-6
Angular Momentum of a Bicycle Wheel	12-6
Angular Velocity as a Vector	12-7
Angular Momentum as a Vector	12-7
Angular Mass or Moment of Inertia	12-7
Calculating Moments of Inertia	12-8
Vector Cross Product	12-9
Right Hand Rule for Cross Products	12-10
Cross Product Definition of Angular Momentum	12-11
The $ec{\mathbf{r}}  imes ec{\mathbf{p}}$ Definition of Angular Momentum	12-12
Angular Analogy to Newton's Second Law	12-14
About Torque	12-15
Conservation of Angular Momentum	12-16
Gvroscopes	12-18
Start-up	12-18
Precession	12-19
Rotational Kinetic Energy	12-22
Combined Translation and Rotation	12-24
Example—Objects Rolling	
Down an Inclined Plane	12-25
Proof of the Kinetic Energy Theorem	12-26

#### CHAPTER 13 EQUILIBRIUM

Equations for equilibrium Example 1 Balancing Weights	13-2 13-2
Gravitational Force acting at the Center of Mass	13-4
Technique of Solving Equilibrium Problems	13-5
Example 3 Wheel and Curb	13-5
Example 4 Rod in a Frictionless Bowl	13-7
Example 5 A Bridge Problem	13-9
Lifting Weights and Muscle Injuries 1	3-11

#### CHAPTER 14 OSCILLATIONS AND RESONANCE

Oscillatory Motion	14-2
The Sine Wave	. 14-3
Phase of an Oscillation	. 14-6
Mass on a Spring;Analytic Solution	14-7
Conservation of Energy	14-11
The Harmonic Oscillator	14-12
The Torsion Pendulum	14-12
The Simple Pendulum	14-15
Small Oscillations	14-16
Simple and Conical Pendulums	14-17
Non Linear Restoring Forces	14-19
Molecular Forces	14-20
Damped Harmonic Motion	14-21
Critical Damping	14-23
Resonance	14-24
Resonance Phenomena	14-26
Transients	14-27
Appendix 14–1 Solution of the Differential Equation	
for Forced Harmonic Motion	14-28
Appendix 14-2 Computer analysis	
of oscillatory motion	14-30
English Program	14-31
The BASIC Program	14-32
Damped Harmonic Motion	14-34

#### CHAPTER 15 ONE DIMENSIONAL WAVE MOTION

Wave Pulses	15-3
Speed of a Wave Pulse	15-4
Dimensional Analysis	15-6
Speed of Sound Waves	15-8
Linear and nonlinear Wave Motion	15-10
The Principle of Superposition	15-11
Sinusoidal Waves	15-12
Wavelength, Period, and Frequency	15-13
Angular Frequency $oldsymbol{\omega}$	15-14
Spacial Frequency k	15-14
Traveling Wave Formula	15-16
Phase and Amplitude	15-17
Standing Waves	15-18
Waves on a Guitar String	15-20
Frequency of Guitar String Waves	15-21
Sound Produced by a Guitar String	15-22

#### CHAPTER 16 FOURIER ANALYSIS, NORMAL MODES AND SOUND

	-
Harmonic Series	16-3
Normal Modes of Oscillation	16-4
Fourier Analysis	16-6
Analysis of a Sine Wave	16-7
Analysis of a Square Wave	16-9
Repeated Wave Forms	16-11
Analysis of the Coupled Air Cart System	16-12
The Human Ear	16-15
Stringed Instruments	16-18
Wind Instruments	16-20
Percussion Instruments	16-22
Sound Intensity	16-24
Bells and Decibels	16-24
Sound Meters	16-26
Speaker Curves	16-27
Appendix A: Fourier Analysis Lecture	16-28
Square Wave	16-28
Calculating Fourier Coefficients	16-28
Amplitude and Phase	16-31
Amplitude and Intensity	16-33
Appendix B: Inside the Cochlea	16-34

#### CHAPTER 17 ATOMS, MOLECULES AND ATOMIC PROCESSES

Molecules 17-2
Atomic Processes
Thermal Motion
Thermal Equilibrium
Temperature 17-9
Absolute Zero 17-9
Temperature Scales 17-10
Molecular Forces 17-12
Evaporation 17-14
Pressure 17-16
Stellar Evolution 17-17
The Ideal Gas Law 17-18
Ideal Gas Thermometer 17-20
The Mercury Barometer
and Pressure Measurements 1/-22
Avogadro's Law 17-24
Heat Capacity 17-26
Specific Heat 17-26
Molar Heat Capacity 17-26
Wolar Specific Heat of Hellum Gas 17-27
Other Gases
Equipartition of Energy 17-28
Federmolecules
Failure of Classical Physics
Thermal Expansion 17.22
Ormatia Expansion 17-33
Usmolic Pressure
Elasticity of Hubber

#### CHAPTER 18 ENTROPY

Introduction	18-2
Work Done by an Expanding Gas	18-5
Specific Heats $C_V$ and $C_p$	18-6
Isothermal Expansion and PV Diagrams	18-8
Isothermal Compression	18-9
Isothermal Expansion of an Ideal Gas	18-9
Adiabatic Expansion	18-9
The Carnot Cycle	18-11
Thermal Efficiency of the Carnot Cycle	18-12
Reversible Engines	18-13
Energy Flow Diagrams	18-15
Maximally Efficient Engines	18-15
Reversibility	18-17
Applications of the Second Law	18-17
Electric Cars	18-19
The Heat Pump	18-19
The Internal Combustion Engine	18-21
Entropy	18-22
The Direction of Time	18-25
Appendix: Calculation of the	
Efficiency of a Carnot Cycle	18-26
Isothermal Expansion	18-26
Adiabatic Expansion	18-26
The Carnot Cycle	18-28

#### CHAPTER 19 THE ELECTRIC INTERACTION

The Four Basic Interactions	. 19-1
Atomic Structure	. 19-3
Isotopes	. 19-6
The Electric Force Law	. 19-7
Strength of the Electric Interaction	. 19-8
Electric Charge	19-8
Positive and Negative Charge	19-10
Addition of Charge	19-10
Conservation of Charge	19-13
Stability of Matter	19-14
Quantization of Electric Charge	19-14
Molecular Forces	19-15
Hydrogen Molecule	19-16
Molecular Forces—A More Quantitative Look	19-18
The Bonding Region	19-19
Electron Binding Energy	19-20
Electron Volt as a Unit of Energy	19-21
Electron Energy in the Hydrogen Molecule Ion	19-21

#### CHAPTER 20 NUCLEAR MATTER

Nuclear Force	20-2
Range of the Nuclear Force	20-3
Nuclear Fission	20-3
Neutrons and the Weak Interaction	20-6
Nuclear Structure	20-7
$oldsymbol{lpha}$ (Alpha) Particles	20-8
Nuclear Binding Energies	20-9
Nuclear Fusion	20-12
Stellar Evolution	20-13
Neutron Stars	20-17
Neutron Stars	
and Black Holes	20-18

#### **CHAPTER 23 FLUID DYNAMICS**

The Current State of Fluid Dynamics	23-1
The Velocity Field	
The Vector Field	23-3
Streamlines	23-4
Continuity Equation	23-5
Velocity Field of a Point Source	23-6
Velocity Field of a Line Source	23-7
Flux	23-8
Bernoulli's Equation	23-9
Applications of Bernoulli's Equation	
Hydrostatics	
Leaky Tank	
Airplane Wing	23-13
Sailboats	23-14
The Venturi Meter	23-15
The Aspirator	23-16
Care in Applying Bernoulli's Equation	23-16
Hydrodynamic Voltage	
Town Water Supply	
Viscous Effects	
Vortices	23-20
Quantized Vortices in Superfluids	

#### CHAPTER 24 COULOMB'S AND GAUSS' LAW

Coulomb's Law	
CGS Units	
MKS Units	
Checking Units in MKS Calculations	
Summary	24-3
Example 1 Two Charges	24-3
Example 2 Hydrogen Atom	24-4
Force Produced by a Line Charge	
Short Rod	24-9
The Electric Field	
Unit Test Charge	
Electric Field lines	
Mapping the Electric Field	
Field Lines	
Continuity Equation for Electric Fields	
Flux	24-15
Negative Charge	24-16
Flux Tubes	
Conserved Field Lines	24-17
A Mapping Convention	
Summary	24-18
A Computer Plot	24-19
Gauss' Law	24-20
Electric Field of a Line Charge	24-21
Flux Calculations	24-22
Area as a Vector	24-22
Gauss' Law for the Gravitational Field	24-23
Gravitational Field of a Point Mass	24-23
Gravitational Field	
of a Spherical Mass	24-24
Gravitational Field Inside the Earth	24-24
Solving Gauss' Law Problems	24-26
Problem Solving	24-29

#### CHAPTER 25 FIELD PLOTS AND ELECTRIC POTENTIAL

The Contour Map	
Equipotential Lines	
Negative and Positive Potential Energy	
Electric Potential of a Point Charge	
Conservative Forces	25-5
Electric Voltage	
A Field Plot Model	
Computer Plots	

#### CHAPTER 26 ELECTRIC FIELDS AND CONDUCTORS

Electric Field Inside a Conductor	26-1
Surface Charges	
Surface Charge Density	
Example: Field in a Hollow Metal Sphere	
Van de Graaff generator	
Electric Discharge	
Grounding	
The Electron Gun	
The Filament	26-9
Accelerating Field	26-10
A Field Plot	26-10
Equipotential Plot	26-11
Electron Volt as a Unit of Energy	26-12
Example	26-13
About Computer Plots	26-13
The Parallel Plate Capacitor	26-14
Deflection Plates	26-16

#### CHAPTER 27 BASIC ELECTRIC CIRCUITS

Electric Current	27- 2
Positive and Negative Currents	27- 3
A Convention	27- 5
Current and Voltage	27- 6
Resistors	27- 6
A Simple Circuit	27- 8
The Short Circuit	27- 9
Power	27- 9
Kirchoff's Law	27- 10
Application of Kirchoff's Law	27- 11
Series Resistors	27- 11
Parallel Resistors	27- 12
Capacitance and Capacitors	27- 14
Hydrodynamic Analogy	27- 14
Cylindrical Tank as a Constant Voltage Source .	27- 15
Electrical Capacitance	27- 16
Energy Storage in Capacitors	27- 18
Energy Density in an Electric Field	27- 19
Capacitors as Circuit Elements	27- 20
The RC Circuit	27- 22
Exponential Decay	27- 23
The Time Constant RC	27- 24
Half-Lives	27- 25
Initial Slope	27- 25
The Exponential Rise	27- 26
The Neon Bulb Oscillator	27- 28
The Neon Bulb	27- 28
The Neon Oscillator Circuit	27- 29
Period of Oscillation	27- 30
Experimental Setup	27- 31

#### CHAPTER 28 MAGNETISM

Two Garden Peas	28- 2
A Thought Experiment	28- 4
Charge Density on the Two Rods	28- 6
A Proposed Experiment	28- 7
Origin of Magnetic Forces	28- 8
Magnetic Forces	28- 10
Magnetic Force Law	28- 10
The Magnetic Field B	28- 10
Direction of the Magnetic Field	28- 11
The Right Hand Rule for Currents	28- 13
Parallel Currents Attract	28- 14
The Magnetic Force Law	28- 14
Lorentz Force Law	28- 15
Dimensions of the	
Magnetic Field, Tesla and Gauss	28- 16
Uniform Magnetic Fields	28- 16
Helmholtz Coils	28- 18
Motion of Charged Particles in Magnetic Field	s 28-19
Motion in a Uniform Magnetic Field	28-20
Particle Accelerators	28-22
Relativistic Energy and Momenta	28- 24
Bubble Chambers	28- 26
The Mass Spectrometer	28- 28
Magnetic Focusing	28- 29
Space Physics	28- 31
The Magnetic Bottle	28- 31
Van Allen Radiation Belts	28- 32

#### CHAPTER 29 AMPERE'S LAW

The Surface Integral	29-2
Gauss' Law	29-3
The Line Integral	29-5
Ampere's Law	29-7
Several Wires	29-10
Field of a Straight Wire	29-11
Field of a Solenoid	29-14
Right Hand Rule for Solenoids	29-14
Evaluation of the Line Integral	29-15
Calculation of <b>i<sub>enclosed</sub></b>	29-15
Using Ampere's law	29-15
One More Right Hand Rule	29-16
The Toroid	29-17

#### CHAPTER 30 FARADAY'S LAW

Electric Field of Static Charges	30-2
A Magnetic Force Experiment	30-3
Air Cart Speed Detector	30-5
A Relativity Experiment	30-9
Faraday's Law	30-11
Magnetic Flux	30-11
One Form of Faraday's Law	30-12
A Circular Electric Field	30-13
Line Integral of <b>Ē</b> around a Closed Path	30-14
Using Faraday's Law	30-15
Electric Field of an Electromagnet	30-15
Right Hand Rule for Faraday's Law	30-15
Electric Field of Static Charges	30-16
The Betatron	30-16
Two Kinds of Fields	30-18
Note on our $\oint \vec{E} \cdot d\vec{\ell}$ meter	30-20
Applications of Faraday's Law	30-21
The AC Voltage Generator	30-21
Gaussmeter	30-23
A Field Mapping Experiment	30-24

#### CHAPTER 31 INDUCTION AND MAGNETIC MOMENT

31-2
31-3
31-4
31-5
31-7
31-8
31-10
31-12
31-13
31-15
31-18
31-18
31-20
31-21
31-22
31-24
31-24
31-26
31-28
31-29
31-30
31-31

#### CHAPTER 32 MAXWELL'S EQUATIONS

Gauss' Law for Magnetic Fields	32	2-2
Maxwell's Correction to Ampere's Law	32	2-4
Example: Magnetic Field		
between the Capacitor Plates	32	- 6
Maxwell's Equations	32	- 8
Symmetry of Maxwell's Equations	32	- 9
Maxwell's Equations in Empty Space	32-	10
A Radiated Electromagnetic Pulse	32-	10
A Thought Experiment	32-	11
Speed of an Electromagnetic Pulse	32-	14
Electromagnetic Waves	32-	18
Electromagnetic Spectrum	32-	20
Components of the Electromagnetic Spectrum.	32-	20
Blackbody Radiation	32-	22
UV, X Rays, and Gamma Rays	32-	22
Polarization	32-	23
Polarizers	32-	24
Magnetic Field Detector	32-	26
Radiated Electric Fields	32-	28
Field of a Point Charge	32-	30

#### CHAPTER 33 LIGHT WAVES

Superposition of	
Circular Wave Patterns	33-2
Huygens Principle	33-4
Two Slit Interference Pattern	33-6
The First Maxima	33-8
Two Slit Pattern for Light	. 33-10
The Diffraction Grating	. 33-12
More About Diffraction Gratings	. 33-14
The Visible Spectrum	. 33-15
Atomic Spectra	. 33-16
The Hydrogen Spectrum	. 33-17
The Experiment on Hydrogen Spectra	. 33-18
The Balmer Series	33-19
The Doppler Effect	. 33-20
Stationary Source and Moving Observer	. 33-21
Doppler Effect for Light	. 33-22
Doppler Effect in Astronomy	. 33-23
The Red Shift and the Expanding Universe	. 33-24
A Closer Look at Interference Patterns	. 33-26
Analysis of the Single Slit Pattern	. 33-27
Recording Diffraction Grating Patterns	. 33-28

#### **CHAPTER 34 PHOTONS**

Blackbody Radiation	34-2
Planck Blackbody Radiation Law	34-4
The Photoelectric Effect	34-5
Planck's Constant h	34-8
Photon Energies	34-9
Particles and Waves	34-11
Photon Mass	34-12
Photon Momentum	34-13
Antimatter	34-16
Interaction of Photons and Gravity	34-18
Evolution of the Universe	34-21
Red Shift and the Expansion of the Universe	34-21
Another View of Blackbody Radiation	34-22
Models of the universe	34-23
Powering the Sun	34-23
Abundance of the Elements	34-24
The Steady State Model of the Universe	34-25
The Big Bang Model	34-26
The Helium Abundance	34-26
Cosmic Radiation	34-27
The Three Degree Radiation	34-27
Thermal Equilibrium of the Universe	34-28
The Early Universe	34-29
The Early Universe	34-29
Excess of Matter over Antimatter	34-29
Decoupling (700,000 years)	34-31
Guidebooks	34-32

#### CHAPTER 35 BOHR THEORY OF HYDROGEN

The Classical Hydrogen Atom	35-2
Energy Levels	35-4
The Bohr Model	35-7
Angular Momentum in the Bohr Model	35-8
De Broglie's Hypothesis	35-10

#### CHAPTER 36 SCATTERING OF WAVES

Scattering of a Wave by a Small Object	36-2
Reflection of Light	36-3
X Ray Diffraction	36-4
Diffraction by Thin Crystals	36-6
The Electron Diffraction Experiment	36-8
The Graphite Crystal	36-8
The Electron Diffraction Tube	36-9
Electron Wavelength	36-9
The Diffraction Pattern	36-10
Analysis of the Diffraction Pattern	36-11
Other Sets of Lines	36-12
Student Projects	36-13
Student project by Gwendylin Chen	36-14

#### CHAPTER 37 LASERS, A MODEL ATOM AND ZERO POINT ENERGY

The Laser and Standing Light Waves	. 37-2
Photon Standing Waves	. 37-3
Photon Energy Levels	. 37-4
A Model Atom	. 37-4
Zero Point Energy	. 37-7
Definition of Temperature	. 37-8
Two dimensional standing waves	. 37-8

#### CHAPTER 38 ATOMS

38-2
38-4
38-5
38-5
38-5
38-7
38-7
38-8
38-9
38-9
38-9
38-10
38-10
38-12
38-12
38-13
38-13
38-13
38-13
38-14
38-14
38-15

#### CHAPTER 39 SPIN

The Concept of Spin 39-3
Interaction of the Magnetic Field with Spin 39-4
Magnetic Moments and the Bohr Magneton 39-4
Insert 2 here
Electron Spin Resonance Experiment
Nuclear Magnetic Moments 39-6
Sign Conventions 39-6
Classical Picture of Magnetic Resonance
Electron Spin Resonance Experiment 39-9
Appendix: Classical Picture of Magnetic Interactions 39-14

#### **CHAPTER 40 QUANTUM MECHANICS**

Two Slit Experiment	. 40-2
The Two Slit Experiment	
from a Particle Point of View	. 40-3
Two Slit Experiment—One Particle at a Time	. 40-3
Born's Interpretation of the Particle Wave	. 40-6
Photon Waves	. 40-6
Reflection and Fluorescence	. 40-8
A Closer Look at the Two Slit Experiment	. 40-9
The Uncertainty Principle	40-14
Position-Momentum Form	
of the Uncertainty Principle	40-15
Single Slit Experiment	40-16
Time-Energy Form of the Uncertainty Principle	40-19
Probability Interpretation	40-22
Measuring Short Times	40-22
Short Lived Elementary Particles	40-23
The Uncertainty Principleand Energy Conservation .	40-24
Quantum Fluctuations and Empty Space	40-25
Appendix: How a pulse is formed from sine waves	40-27

#### **CHAPTER ON GEOMETRICAL OPTICS**

Reflection from Curved Surfaces	Optics-3
The Parabolic Reflection	Optics-4
Mirror Images	Optics-6
The Corner Reflector	Optics-7
Motion of Light through a Medium	Optics-8
Index of Refraction	Optics-9
Cerenkov Radiation	. Optics-10
Snell's Law	. Optics-11
Derivation of Snell's Law	. Optics-12
Internal Reflection	. Optics-13
Fiber Optics	. Optics-14
Medical Imaging	. Optics-15
Prisms	. Optics-15
Rainbows	. Optics-16
The Green Flash	. Optics-17
Halos and Sun Dogs	. Optics-18
Lenses	. Optics-18
Spherical Lens Surface	. Optics-19
Focal Length of a Spherical Surface	. Optics-20
Aberrations	. Optics-21
Thin Lenses	. Optics-23
The Lens Equation	. Optics-24
Negative Image Distance	. Optics-26
Negative Focal Length & Diverging Lenses	. Oplics-26
Multiple Long Systems	Optics-27
Two Longos Togothor	Optice 20
Magnification	Optice-20
The Human Eve	Optice 31
Nearsightedness and Farsightedness	Ontics-32
The Camera	Option 22
Depth of Field	Ontics-34
Eve Glasses and a Home Lab Experiment	Ontics-36
The Eveniece	Ontice-37
The Magnifier	Ontics-38
Angular Magnification	Optics-39
Telescones	Ontics-40
Reflecting telescopes	Optics-42
Large Reflecting Telescopes.	. Optics-43
Hubbel Space Telescope	. Optics-44
World's Largest Optical Telescope	. Optics-45
Infrared Telescopes	. Optics-46
Radio Telescopes	. Optics-48
The Very Long Baseline Array (VLBA)	. Optics-49
Microscopes	. Optics-50
Scanning Tunneling Microscope	. Optics-51
Photograph credits	i

### A PHYSICS BASED CALCULUS TEXT

#### CHAPTER 1 INTRODUCTION TO CALCULUS

Limiting Process	Cal 1-3
The Uncertainty Principle	Cal 1-3
Calculus Definition of Velocity	Cal 1-5
Acceleration	Cal 1-7
Components	Cal 1-7
Integration	Cal 1-8
Prediction of Motion	Cal 1-9
Calculating Integrals	Cal 1-11
The Process of Integrating	Cal 1-13
Indefinite Integrals	Cal 1-14
Integration Formulas	Cal 1-14
New Functions	Cal 1-15
Logarithms	Cal 1-15
The Exponential Function	Cal 1-16
Exponents to the Base 10	Cal 1-16
The Exponential Function <b>y<sup>x</sup></b>	Cal 1-16
Euler's Number e = 2.7183	Cal 1-17
Differentiation and Integration	Cal 1-18
A Fast Way to go Back and Forth	Cal 1-20
Constant Acceleration Formulas	Cal 1-20
Constant Acceleration Formulas	
in Three Dimensions	Cal 1-22
More on Differentiation	Cal 1-23
Series Expansions	Cal 1-23
Derivative of the Function $\mathbf{x}^{\mathbf{n}}$	Cal 1-24
The Chain Rule	Cal 1-25
Remembering The Chain Rule	Cal 1-25
Partial Proof of the Chain Rule (optional).	Cal 1-26
Integration Formulas	Cal 1-27
Derivative of the Exponential Function	Cal 1-28
Integral of the Exponential Function	Cal 1-29
Derivative as the Slope of a Curve	Cal 1-30
Negative Slope	Cal 1-31
The Exponential Decay	Cal 1-32
Muon Lifetime	Cal 1-32
Half Life	Cal 1-33
Measuring the Time	
Constant from a Graph	Cal 1-34
The Sine and Cosine Functions	Cal 1-35
Radian Measure	Cal 1-35
The Sine Function	Cal 1-36
Amplitude of a Sine Wave	Cal 1-37
Derivative of the Sine Function	Cal 1-38
Physical Constants in CGS Units	Back cover-1
Conversion Factors	Back cover-1

# Physics 2000

## E. R. Huggins **Dartmouth College**

## Part I Mechanics, **Waves & Particles**



physics2000.com

## Introduction An Overview of Physics

With a brass tube and a few pieces of glass, you can construct either a microscope or a telescope. The difference is essentially where you place the lenses. With the microscope, you look down into the world of the small, with the telescope out into the world of the large.

In the twentieth century, physicists and astronomers have constructed ever larger machines to study matter on even smaller or even larger scales of distance. For the physicists, the new microscopes are the particle accelerators that provide views well inside atomic nuclei. For the astronomers, the machines are radio and optical telescopes whose large size allows them to record the faintest signals from space. Particularly effective is the Hubble telescope that sits above the obscuring curtain of the earth's atmosphere.

The new machines do not provide a direct image like the ones you see through brass microscopes or telescopes. Instead a good analogy is to the Magnetic Resonance Imaging (MRI) machines that first collect a huge amount of data, and then through the use of a computer program construct the amazing images showing cross sections through the human body. The telescopes and particle accelerators collect the vast amounts of data. Then through the use of the theories of quantum mechanics and relativity, the data is put together to construct meaningful images.

Some of the images have been surprising. One of the greatest surprises is the increasingly clear image of the universe starting out about fourteen billion years ago

as an incredibly small, incredibly hot speck that has expanded to the universe we see today. By looking farther and farther out, astronomers have been looking farther and farther back in time, closer to that hot, dense beginning. Physicists, by looking at matter on a smaller and smaller scale with the even more powerful accelerators, have been studying matter that is even hotter and more dense. By the end of the twentieth century, physicists and astronomers have discovered that they are looking at the same image.

It is likely that telescopes will end up being the most powerful microscopes. There is a limit, both financial and physical, to how big and powerful an accelerator we can build. Because of this limit, we can use accelerators to study matter only up to a certain temperature and density. To study matter that is still hotter and more dense, which is the same as looking at still smaller scales of distance, the only "machine" we have available is the universe itself. We have found that the behavior of matter under the extreme conditions of the very early universe have left an imprint that we can study today with telescopes.

In the rest of this introduction we will show you some of the pictures that have resulted from looking at matter with the new machines. In the text itself we will begin to learn how these pictures were constructed.

#### SPACE AND TIME

The images of nature we see are images in both space and time, for we have learned from the work of Einstein that the two cannot be separated. They are connected by the speed of light, a quantity we designate by the letter c, which has the value of a billion (1,000,000,000) feet (30 cm) in a second. Einstein's remarkable discovery in 1905 was that the speed of light is an absolute speed limit. Nothing in the current universe can travel faster than the speed c.

Because the speed of light provides us with an absolute standard that can be measured accurately, we use the value of c to relate the definitions of time and distance. The meter is defined as the distance light travels in an interval of 1/299,792.458 of a second. The length of a second itself is provided by an atomic standard. It is the time interval occupied by 9,192,631,770 vibrations of a particular wavelength of light radiated by a cesium atom.

Using the speed of light for conversion, clocks often make good meter sticks, especially for measuring astronomical distances. It takes light 1.27 seconds to travel from the earth to the moon. We can thus say that the moon is 1.27 *light seconds* away. This is simpler than saying that the moon is 1,250,000,000 feet or 382,000 kilometers away. Light takes 8 minutes to reach us from the sun, thus the earth's orbit about the sun has a radius of 8 *light minutes*. Radio signals, which also travel at the speed of light, took 2 1/2 hours to reach the earth when Voyager II passed the planet Uranus (temporarily the most distant planet). Thus Uranus is 2 1/2 light hours away and our solar system



**Figure 1** *The Sombrero galaxy.* 

has a diameter of 5 light hours (not including the cloud of comets that lie out beyond the planets.)

The closest star, Proxima Centauri, is 4.2 *light years* away. Light from this star, which started out when you entered college as a freshman, will arrive at the earth shortly after you graduate (assuming all goes well). Stars in our local area are typically 2 to 4 light years apart, except for the so called *binary stars* which are pairs of stars orbiting each other at distances as small as light days or light hours.

On a still larger scale, we find that stars form island structures called *galaxies*. We live in a fairly typical galaxy called the Milky Way. It is a flat disk of stars with a slight bulge at the center much like the Sombrero Galaxy seen edge on in Figure (1) and the neighboring spiral galaxy Andromeda seen in Figure (2). Our Milky Way is a spiral galaxy much like Andromeda, with the sun located about 2/3 of the way out in one of the spiral arms. If you look at the sky on a dark clear night you can see the band of stars that cross the sky called the Milky Way. Looking at these stars you are looking sideways through the disk of the Milky Way galaxy.



**Figure 2** *The Andromeda galaxy*.

Our galaxy and the closest similar galaxy, Andromeda, are both about 100,000 light years (.1 million light years) in diameter, contain about a billion stars, and are about one million light years apart. These are more or less typical numbers for the average size, population and spacing of galaxies in the universe.

To look at the universe over still larger distances, first imagine that you are aboard a rocket leaving the earth at night. As you leave the launch pad, you see the individual lights around the launch pad and street lights in neighboring roads. Higher up you start to see the lights from the neighboring city. Still higher you see the lights from a number of cities and it becomes harder and harder to see individual street lights. A short while later all the bright spots you see are cities, and you can no longer see individual lights. At this altitude you count cities instead of light bulbs.

Similarly on our trip out to larger and larger distances in the universe, the bright spots are the galaxies for we can no longer see the individual stars inside. On distances ranging from millions up to billions of light years, we see galaxies populating the universe. On this scale they are small but not quite point like. Instruments like the Hubble telescope in space can view structure in the most distant galaxies, like those shown in Figure (3).



**Figure 3** *Hubble photograph of the most distant galaxies.* 

#### The Expanding Universe

In the 1920s, Edwin Hubble made the surprising discovery that, on average, the galaxies are all moving away from us. The farther away a galaxy is, the faster it is moving away. Hubble found a simple rule for this recession, a galaxy twice as far away is receding twice as fast.

At first you might think that we are at the exact center of the universe if the galaxies are all moving directly away from us. But that is not the case. Hubble's discovery indicates that the universe is expanding uniformly. You can see how a uniform expansion works by blowing up a balloon part way, and drawing a number of uniformly spaced dots on the balloon. Then pick any dot as your own dot, and watch it as you continue to blow the balloon up. You will see that the neighboring dots all move away from your dot, and you will also observe Hubble's rule that dots twice as far away move away twice as fast.

Hubble's discovery provided the first indication that there is a limit to how far away we can see things. At distances of about fourteen billion light years, the recessional speed approaches the speed of light. Recent photographs taken by the Hubble telescope show galaxies receding at speeds in excess of 95% the speed of light, galaxies close to the edge of what we call the *visible universe*.

The implications of Hubble's rule are more dramatic if you imagine that you take a moving picture of the expanding universe and then run the movie backward in time. The rule that galaxies twice as far away are receding twice as fast become the rule that galaxies twice as far away are approaching you twice as fast. A more distant galaxy, one at twice the distance but heading toward you at twice the speed, will get to you at the same time as a closer galaxy. In fact, all the galaxies will reach you at the same instant of time.

Now run the movie forward from that instant of time, and you see all the galaxies flying apart from what looks like a single explosion. From Hubble's law you can figure that the explosion should have occurred about fourteen billion years ago. Did such an explosion really happen, or are we simply misreading the data? Is there some other way of interpreting the expansion without invoking such a cataclysmic beginning? Various astronomers thought there was. In their *continuous creation theory* they developed a model of the universe that was both unchanging and expanding at the same time. That sounds like an impossible trick because as the universe expands and the galaxies move apart, the density of matter has to decrease. To keep the universe from changing, the model assumed that matter was being created throughout space at just the right rate to keep the average density of matter constant.

With this theory one is faced with the question of which is harder to accept—the picture of the universe starting in an explosion which was derisively called the **Big Bang**, or the idea that matter is continuously being created everywhere? To provide an explicit test of the continuous creation model, it was proposed that all matter was created in the form of hydrogen atoms, and that all the elements we see around us today, the carbon, oxygen, iron, uranium, etc., were made as a result of nuclear reactions inside of stars.

To test this hypothesis, physicists studied in the laboratory those nuclear reactions which should be relevant to the synthesis of the elements. The results were quite successful. They predicted the correct or nearly correct abundance of all the elements but one. The holdout was helium. There appeared to be more helium in the universe than they could explain. By 1960, it was recognized that, to explain the abundance of the elements as a result of nuclear reactions inside of stars, you have to start with a mixture of hydrogen and helium. Where did the helium come from? Could it have been created in a Big Bang?

As early as 1948, the Russian physicist George Gamov studied the consequences of the Big Bang model of the universe. He found that if the conditions in the early universe were just right, there should be light left over from the explosion, light that would now be a faint glow at radio wave frequencies. Gamov talked about this prediction with several experimental physicists and was told that the glow would be undetectable. Gamov's prediction was more or less ignored until 1964 when the glow was accidently detected as noise in a radio telescope. Satellites have now been used to study this glow in detail, and the results leave little doubt about the explosive nature of the birth of the universe.

What was the universe like at the beginning? In an attempt to find out, physicists have applied the laws of physics, as we have learned them here on earth, to the collapsing universe seen in the time reversed motion picture of the galaxies. One of the main features that emerges as we go back in time and the universe gets smaller and smaller, is that it also becomes hotter and hotter. The obvious question in constructing a model of the universe is how small and how hot do we allow it to get? Do we stop our model, stop our calculations, when the universe is down to the size of a galaxy? a star? a grapefruit? or a proton? Does it make any sense to apply the laws of physics to something as hot and dense as the universe condensed into something smaller than, say, the size of a grapefruit? Surprisingly, it may. One of the frontiers of physics research is to test the application of the laws of physics to this model of the hot early universe.

We will start our disruption of the early universe at a time when the universe was about a billionth of a second old and the temperature was three hundred thousand billion  $(3 \times 10^{14})$  degrees. While this sounds like a preposterously short time and unbelievably high temperature, it is not the shortest time or highest temperature that has been quite carefully considered. For our overview, we are arbitrarily choosing that time because of the series of pictures we can paint which show the universe evolving. These pictures all involve the behavior of matter as it has been studied in the laboratory. To go back earlier relies on theories that we are still formulating and trying to test.

To recognize what we see in this evolving picture of the universe, we first need a reasonably good picture of what the matter around us is like. With an understanding of the building blocks of matter, we can watch the pieces fit together as the universe evolves. Our discussion of these building blocks will begin with atoms which appear only late in the universe, and work down to smaller particles which play a role at earlier times. To understand what is happening, we also need a picture of how matter interacts via the basic forces in nature.

When you look through a microscope and change the magnification, what you see and how you interpret it, changes, even though you are looking at the same sample. To get a preliminary idea of what matter is made from and how it behaves, we will select a particular sample and magnify it in stages. At each stage we will provide a brief discussion to help interpret what we see. As we increase the magnification, the interpretation of what we see changes to fit and to explain the new picture. Surprisingly, when we get down to the smallest scales of distance using the greatest magnification, we see the entire universe at its infancy. We have reached the point where studying matter on the very smallest scale requires an understanding of the very largest, and vice versa.

#### STRUCTURE OF MATTER

We will start our trip down to small scales with a rather large, familiar example—the earth in orbit about the sun. The earth is attracted to the sun by a force called *gravity*, and its motion can be accurately forecast, using a set of rules called *Newtonian mechanics*. The basic concepts involved in Newtonian mechanics are force, mass, velocity and acceleration, and the rules tell us how these concepts are related. (Half of the traditional introductory physics courses is devoted to learning these rules.)

#### Atoms

We will avoid much of the complexity we see around us by next focusing in on a single hydrogen atom. If we increase the magnification so that a garden pea looks as big as the earth, then one of the hydrogen atoms inside the pea would be about the size of a basketball. How we interpret what we see inside the atom depends upon our previous experience with physics. With a background in Newtonian mechanics, we would see a miniature solar system with the nucleus at the center and an electron in orbit. The nucleus in hydrogen consists of a single particle called the proton, and the electron is held in orbit by an electric force. At this magnification, the proton and electron are tiny points, too small to show any detail.



Elliptical orbit of an earth satellite calculated using Newtonian mechanics.

There are similarities and striking differences between the gravitational force that holds our solar system together and the electric force that holds the hydrogen atom together. Both forces in these two examples are attractive, and both forces decrease as the square of the distance between the particles. That means that if you double the separation, the force is only one quarter as strong. The strength of the gravitational force depends on the *mass* of the objects, while the electric force depends upon the *charge* of the objects.

One of the major differences between electricity and gravity is that all gravitational forces are attractive, while there are both attractive and repulsive electric forces. To account for the two types of electric force, we say that there are two kinds of electric charge, which Benjamin Franklin called *positive charge* and *negative charge*. The rule is that like charges repel while opposite charges attract. Since the electron and the proton have opposite charge they attract each other. If you tried to put two electrons together, they would repel because they have like charges. You get the same repulsion between two protons. By the accident of Benjamin Franklin's choice, protons are positively charged and electrons are negatively charged.

On a large scale, gravity dominates because of the cancellation of electric forces. Consider, for example, the net electric force between two complete hydrogen atoms separated by some small distance. Call them atom A and atom B. Between these two atoms there are four distinct forces, two attractive and two repulsive. The attractive forces are between the proton in atom A and the electron in atom B, and between the electron in atom A and the proton in atom B. However, the two

protons repel each other and the electrons repel to give the two repulsive forces. The net result is that the attractive and repulsive forces cancel and we end up with essentially no electric force between the atoms.

Rather than counting individual forces, it is easier to add up electric charge. Since a proton and an electron have opposite charges, the total charge in a hydrogen atom adds up to zero. With no net charge on either of the two hydrogen atoms in our example, there is no net electric force between them. We say that a complete hydrogen atom is *electrically neutral*.

While complete hydrogen atoms are neutral, they can attract each other if you bring them too close together. What happens is that the electron orbits are distorted by the presence of the neighboring atom, the electric forces no longer exactly cancel, and we are left with a small residual force called a *molecular force*. It is the molecular force that can bind the two hydrogen atoms together to form a hydrogen molecule. These molecular forces are capable of building very complex objects, like people. We are the kind of structure that results from electric forces, in much the same way that solar systems and galaxies are the kind of structures that result from gravitational forces.

Chemistry deals with reactions between about 100 different elements, and each element is made out of a different kind of atom. The basic distinction between atoms of different elements is the number of protons in the nucleus. A hydrogen nucleus has one proton, a helium nucleus 2 protons, a lithium nucleus 3 protons, on up to the largest naturally occurring nucleus, uranium with 92 protons.

Complete atoms are electrically neutral, having as many electrons orbiting outside as there are protons in the nucleus. The chemical properties of an atom are determined almost exclusively by the structure of the orbiting electrons, and their electron structure depends very much on the number of electrons. For example, helium with 2 electrons is an inert gas often breathed by deep sea divers. Lithium with 3 electrons is a reactive metal that bursts into flame when exposed to air. We go from an inert gas to a reactive metal by adding one electron.

#### Light

The view of the hydrogen atom as a miniature solar system, a view of the atom seen through the "lens" of Newtonian mechanics, fails to explain much of the atom's behavior. When you heat hydrogen gas, it glows with a reddish glow that consists of three distinct colors or so called *spectral lines*. The colors of the lines are bright red, swimming pool blue, and deep violet. You need more than Newtonian mechanics to understand why hydrogen emits light, let alone explain these three special colors.

In the middle of the 1800s, Michael Faraday went a long way in explaining electric and magnetic phenomena in terms of *electric* and *magnetic fields*. These fields are essentially maps of electric and magnetic forces. In 1860 James Clerk Maxwell discovered that the four equations governing the behavior of electric and magnetic fields could be combined to make up what is called a *wave equation*. Maxwell could construct his wave equation after making a small but crucial correction to one of the underlying equations.

The importance of Maxwell's wave equation was that it predicted that a particular combination of electric and magnetic fields could travel through space in a wavelike manner. Equally important was the fact that the wave equation allowed Maxwell to calculate what the speed of the wave should be, and the answer was about a billion feet per second. Since only light was known to travel that fast, Maxwell made the guess that he had discovered the theory of light, that *light consisted of a wave of electric and magnetic fields of force*. Visible light is only a small part of what we call the *electromagnetic spectrum*. Our eyes are sensitive to light waves whose wavelength varies only over a very narrow range. Shorter wavelengths lie in the *ultravio-let* or *x ray* region, while at increasingly longer wavelengths are *infra red light, microwaves*, and *radio waves*. Maxwell's theory made it clear that these other wavelengths should exist, and within a few years, radio waves were discovered. The broadcast industry is now dependent on Maxwell's equations for the design of radio and television transmitters and receivers. (Maxwell's theory is what is usually taught in the second half of an introductory physics course. That gets you all the way up to 1860.)

While Maxwell's theory works well for the design of radio antennas, it does not do well in explaining the behavior of a hydrogen atom. When we apply Maxwell's theory to the miniature solar system model of hydrogen, we do predict that the orbiting electron will radiate light. But we also predict that the atom will self destruct. The unambiguous prediction is that the electron will continue to radiate light of shorter and shorter wavelength while spiraling in faster and faster toward the nucleus, until it crashes. The combination of Newton's laws and Maxwell's theory is known as *Classical Physics*. We can easily see that classical physics fails when applied even to the simplest of atoms.



The electromagnetic spectrum.

#### Photons

In the late 1890's, it was discovered that a beam of light could knock electrons out of a hydrogen atom. The phenomenon became known as the photoelectric effect. You can use Maxwell's theory to get a rough idea of why a wave of electric and magnetic force might be able to pull electrons out of a surface, but the details all come out wrong. In 1905, in the same year that he developed his theory of relativity, Einstein explained the photoelectric effect by proposing that light consisted of a beam of particles we now call *photons*. When a metal surface is struck by a beam of photons, an electron can be knocked out of the surface if it is struck by an individual photon. A simple formula for the energy of the photons led to an accurate explanation of all the experimental results related to the photoelectric effect.

Despite its success in explaining the photoelectric effect, Einstein's photon picture of light was in conflict not only with Maxwell's theory, it conflicted with over 100 years of experiments which had conclusively demonstrated that light was a wave. This conflict was not to be resolved in any satisfactory way until the middle 1920s.

The particle nature of light helps but does not solve the problems we have encountered in understanding the behavior of the electron in hydrogen. According to Einstein's photoelectric formula, the energy of a photon is inversely proportional to its wavelength. The longer wavelength red photons have less energy than the shorter wavelength blue ones. To explain the special colors of light emitted by hydrogen, we have to be able to explain why only photons with very special energies can be emitted.

#### The Bohr Model

In 1913, the year after the nucleus was discovered, Neils Bohr developed a somewhat ad hoc model that worked surprisingly well in explaining hydrogen. Bohr assumed that the electron in hydrogen could travel on only certain *allowed orbits*. There was a smallest, lowest energy orbit that is occupied by an electron in cool hydrogen atoms. The fact that this was the *smallest* allowed orbit meant that the electron would not spiral in and crush into the nucleus.

Using Maxwell's theory, one views the electron as radiating light continuously as it goes around the orbit. In Bohr's picture the electron does not radiate while in one of the allowed orbits. Instead it radiates, it emits a photon, only when it jumps from one orbit to another.

To see why heated hydrogen radiates light, we need a picture of thermal energy. A gas, like a bottle of hydrogen or the air around us, consists of molecules flying around, bouncing into each other. Any moving object has extra energy due to its motion. If all the parts of the object are moving together, like a car traveling down the highway, then we call this energy of motion *kinetic energy*. If the motion is the random motion of molecules bouncing into each other, we call it *thermal energy*.

The temperature of a gas is proportional to the average thermal energy of the gas molecules. As you heat a gas, the molecules move faster, and their average thermal



**Figure 35-6** *The allowed orbits of the Bohr Model.* 

energy and temperature rises. At the increased speed the collisions between molecules are also stronger.

Consider what happens if we heat a bottle of hydrogen gas. At room temperature, before we start heating, the electrons in all the atoms are sitting in their lowest energy orbits. Even at this temperature the atoms are colliding but the energy involved in a room temperature collision is not great enough to knock an electron into one of the higher energy orbits. As a result, room temperature hydrogen does not emit light.

When you heat the hydrogen, the collisions between atoms become stronger. Finally you reach a temperature in which enough energy is involved in a collision to knock an electron into one of the higher energy orbits. The electron then falls back down, from one allowed orbit to another until it reaches the bottom, lowest energy orbit. The energy that the electron loses in each fall, is carried out by a photon. Since there are only certain allowed orbits, there are only certain special amounts of energy that the photon can carry out.

To get a better feeling for how the model works, suppose we number the orbits, starting at orbit 1 for the lowest energy orbit, orbit 2 for the next lowest energy orbit, etc. Then it turns out that the photons in the red spectral line are radiated when the electron falls from orbit 3 to orbit 2. The red photon's energy is just equal to the energy the electron loses in falling between these orbits. The more energetic blue photons carry out the energy an electron loses in falling from orbit 4 to orbit 2, and the still more energetic violet photons correspond to a fall from orbit 5 to orbit 2. All the other jumps give rise to photons whose energy is too large or too small to be visible. Those with too much energy are ultraviolet photons, while those with too little are in the infra red part of the spectrum. The jump down to orbit 1 is the biggest jump with the result that all jumps down to the lowest energy orbit results in ultraviolet photons.

It appears rather ad hoc to propose a theory where you invent a large number of special orbits to explain what we now know as a large number of spectral lines. One criterion for a successful theory in science is that you get more out of the theory than you put in. If Bohr had to invent a new allowed orbit for each spectral line explained, the theory would be essentially worthless. However this is not the case for the Bohr model. Bohr found a simple formula for the electron energies of all the allowed orbits. This one formula in a sense explains the many spectral lines of hydrogen. A lot more came out of Bohr's model than Bohr had to put in.

The problem with Bohr's model is that it is essentially based on Newtonian mechanics, but there is no excuse whatsoever in Newtonian mechanics for identifying any orbit as special. Bohr focused the problem by discovering that the allowed orbits had special values of a quantity called *angular momentum*.

Angular momentum is related to rotational motion, and in Newtonian mechanics angular momentum increases continuously and smoothly as you start to spin an object. Bohr could explain his allowed orbits by proposing that there was a special unique value of angular momentum—call it a *unit of angular momentum*. Bohr found, using standard Newtonian calculations, that his lowest energy orbit had one unit of angular momentum, orbit 2 had two units, orbit 3 three units, etc. Bohr could explain his entire model by the one assumption that angular momentum was *quantized*, i.e., came only in units.

Bohr's quantization of angular momentum is counter intuitive, for it leads to the picture that when we start to rotate an object, the rotation increases in a jerky fashion rather than continuously. First the object has no angular momentum, then one unit, then 2 units, and on up. The reason we do not see this jerky motion when we start to rotate something large like a bicycle wheel, is that the basic unit of angular momentum is very small. We cannot detect the individual steps in angular momentum, it seems continuous. But on the scale of an atom, the steps are big and have a profound effect.

With Bohr's theory of hydrogen and Einstein's theory of the photoelectric effect, it was clear that classical physics was in deep trouble. Einstein's photons gave a lumpiness to what should have been a smooth wave in Maxwell's theory of light and Bohr's model gave a jerkiness to what should be a smooth change in angular momentum. The bumps and jerkiness needed a new picture of the way matter behaves, a picture that was introduced in 1924 by the graduate student Louis de Broglie.

## PARTICLE-WAVE NATURE OF MATTER

Noting the wave and particle nature of light, de Broglie proposed that the electron had both a wave and a particle nature. While electrons had clearly exhibited a particle behavior in various experiments, de Broglie suggested that it was the wave nature of the electron that was responsible for the special allowed orbits in Bohr's theory. De Broglie presented a simple wave picture where, in the allowed orbits, an integer number of wavelengths fit around the orbit. Orbit 1 had one wavelength, orbit 2 had two wavelengths, etc. In De Broglie's picture, electron waves in non allowed orbits would cancel themselves out. Borrowing some features of Einstein's photon theory of light waves, de Broglie could show that the angular momentum of the electron would have the special quantized values when the electron wave was in one of the special, non cancelling orbits.

With his simple wave picture, de Broglie had hit upon the fundamental idea that was missing in classical physics. The idea is that *all matter*, not just light, has a *particle wave nature*.

It took a few years to gain a satisfactory interpretation of the dual particle wave nature of matter. The current interpretation is that things like photons are in fact particles, but their motion is governed, not by Newtonian mechanics, but by the laws of wave motion. How this works in detail is the subject of our chapter on Quantum Mechanics. One fundamental requirement of our modern interpretation of the particle wave is that, for the interpretation to be meaningful, all forms of matter, without exception, must have this particle wave nature. This general requirement is summarized by a rule discovered by Werner Heisinberg, a rule known as the *uncertainty principle*. How the rule got that name is also discussed in our chapter on quantum mechanics.

In 1925, after giving a seminar describing de Broglie's model of electron waves in hydrogen, Erwin Schrödinger was chided for presenting such a "child-ish" model. A colleague reminded him that waves do not work that way, and suggested that since Schrödinger had nothing better to do, he should work out a real wave equation for the electron waves, and present the results in a couple of weeks.

It took Schrödinger longer than a couple of weeks, but he did succeed in constructing a wave equation for the electron. In many ways Schrödinger's wave equation for the electron is analogous to Maxwell's wave equation for light. Schrödinger's wave equation for the electron allows one to calculate the behavior of electrons in all kinds of atoms. It allows one to explain and predict an atom's electron structure and chemical properties. Schrödinger's equation has become the fundamental equation of chemistry.



**Figure 35-9** De Broglie picture of an electron wave cancelling itself out.



**Figure 35-10** If the circumference of the orbit is an integer number of wavelengths, the electron wave will go around without any cancellation.

#### **CONSERVATION OF ENERGY**

Before we go on with our investigation of the hydrogen atom, we will take a short break to discuss the idea of conservation of energy. This idea, which originated in Newtonian mechanics, survives more or less intact in our modern particle-wave picture of matter.

Physicists pay attention to the concept of energy only because energy is conserved. If energy disappears from one place, it will show up in another. We saw this in the Bohr model of hydrogen. When the electron lost energy falling down from one allowed orbit to a lower energy orbit, the energy lost by the electron was carried out by a photon.

You can store energy in an object by doing *work* on the object. When you lift a ball off the floor, for example, the work you did lifting the ball, the energy you supplied, is stored in a form we call *gravitational potential energy*. Let go of the ball and it falls to the floor, loosing its gravitational potential energy. But just before it hits the floor, it has a lot of energy of motion, what we have called kinetic energy. All the gravitational potential energy the ball had before we dropped it has been converted to kinetic energy.

After the ball hits the floor and is finally resting there, it is hard to see where the energy has gone. One place it has gone is into thermal energy, the floor and the ball are a tiny bit warmer as a result of your dropping the ball.

Another way to store energy is to compress a spring. When you release the spring you can get the energy back. For example, compress a watch spring by winding up the watch, and the energy released as the spring unwinds will run the watch for a day. We could call the energy stored in the compressed spring *spring potential energy*. Physicists invent all sorts of names for the various forms of energy. One of the big surprises in physics was Einstein's discovery of the equivalence of mass and energy, a relationship expressed by the famous equation  $E = mc^2$ . In that equation, E stands for the energy of an object, m its mass, and c is the speed of light. Since the factor  $c^2$  is a constant, Einstein's equation is basically saying that mass is a form of energy. The  $c^2$  is there because mass and energy were initially thought to be different quantities with different units like kilograms and joules. The  $c^2$  simply converts mass units into energy units.

What is amazing is the amount of energy that is in the form of mass. If you could convert all the mass of a pencil eraser into electrical energy, and sell the electrical energy at the going rate of 10¢ per kilowatt hour, you would get about 10 million dollars for it. The problem is converting the mass to another, more useful, form of energy. If you can do the conversion, however, the results can be spectacular or terrible. Atomic and hydrogen bombs get their power from the conversion of a small fraction of their mass energy into thermal energy. The sun gets its energy by "burning" hydrogen nuclei to form helium nuclei. The energy comes from the fact that a helium nucleus has slightly less mass than the hydrogen nuclei out of which it was formed.

If you have a particle at rest and start it moving, the particle gains kinetic energy. In Einstein's view the particle at rest has energy due to its *rest mass*. When you start the particle moving, it gains energy, and since mass is equivalent to energy, it also gains mass. For most familiar speeds the increase in mass due to kinetic energy is very small. Even at the speeds travelled by rockets and spacecraft, the increase in mass due to kinetic energy is hardly noticeable. Only when a particle's speed gets up near the speed of light does the increase in mass become significant. One of the first things we discussed about the behavior of matter is that nothing can travel faster than the speed of light. You might have wondered if nature had traffic cops to enforce this speed limit. It does not need one, it uses a law of nature instead. As the speed of an object approaches the speed of light, its mass increases. The closer to the speed of light, the greater increase in mass. To push a particle up to the speed of light would give it an infinite mass and therefore require an infinite amount of energy. Since that much energy is not available, no particle is going to exceed nature's speed limit.

This raises one question. What about photons? They are particles of light and therefore travel *at* the speed of light. But their energy is not infinite. It depends instead on the wavelength or color of the photon. Photons escape the rule about mass increasing with speed by starting out with no rest mass. You stop a photon and nothing is left. Photons can only exist by traveling at the speed of light.

When a particle is traveling at speeds close enough to the speed of light that its kinetic energy approaches its rest mass energy, the particle behaves differently than slowly moving particles. For example, push on a slowly moving particle and you can make the particle move faster. Push on a particle already moving at nearly the speed of light, and you merely make the particle more massive since it cannot move faster. Since the relationship between mass and energy came out of Einstein's theory of relativity, we say that particles moving near the speed of light obey *relativistic* mechanics while those moving slowly are *nonrelativistic*. Light is always relativistic, and all automobiles on the earth are nonrelativistic.

#### **ANTI-MATTER**

Schrödinger's equation for electron waves is a nonrelativistic theory. It accurately describes electrons that are moving at speeds small compared to the speed of light. This is fine for most studies in chemistry, where chemical energies are much much less than rest mass energies. You can see the difference for example by comparing the energy released by a conventional chemical bomb and an atomic bomb.

Schrödinger of course knew Einstein's theory of relativity, and initially set out to derive a *relativistic wave equation* for the electron. This would be an equation that would correctly explain the behavior of electrons even as the speed of the electrons approached the speed of light and their kinetic energy became comparable to or even exceeded their rest mass energy.

Schrödinger did construct a relativistic wave equation. The problem was that the equation had two solutions, one representing ordinary electrons, the other an apparently impossible particle with a negative rest mass. In physics and mathematics we are often faced with equations with two or more solutions. For example, the formula for the hypotenuse c of a right triangle with sides of lengths a and b is

$$c^2 = a^2 + b^2$$

This equation has two solutions, namely  $c = +\sqrt{a^2 + b^2}$  and  $c = -\sqrt{a^2 + b^2}$ . The negative solution does not give us much of a problem, we simply ignore it.

Schrödinger could not ignore the negative mass solutions in his relativistic wave equation for the following reason. If he started with just ordinary positive mass electrons and let them interact, the equation predicted that the negative mass solutions would be created! The peculiar solutions could not be ignored if the equation was to be believed. Only by going to his nonrelativistic equation could Schrödinger avoid the peculiar solutions.

Int-13

A couple years later, Dirac tried again to develop a relativistic wave equation for the electron. At first it appeared that Dirac's equation would avoid the negative mass solutions, but with little further work, Dirac found that the negative mass solutions were still there. Rather than giving up on his new equation, Dirac found a new interpretation of these peculiar solutions. Instead of viewing them as negatively charged electrons with a negative mass, he could interpret them as positive mass particles with a positive electric charge. According to Dirac's equation, positive and negative charged solutions could be created or destroyed in pairs. The pairs could be created any time enough energy was available.

Dirac predicted the existence of this positively charged particle in 1929. It was not until 1933 that Carl Anderson at Caltech, who was studying the elementary particles that showered down from the sky (particles called *cosmic rays*), observed a positively charged particle whose mass was the same as that of the electron. Named the *positron*, this particle was immediately identified as the positive particle expected from Dirac's equation.

In our current view of matter, all particles are described by relativistic wave equations, and all relativistic wave equations have two kinds of solutions. One solution is for ordinary matter particles like electrons, protons, and neutrons. The other solution, which we now call *antimatter*, describes anti particles, the *antielectron* which is the positron, and the *antiproton* and the *antineutron*. Since all antiparticles can be created or destroyed in particle-anti particle pairs, the antiparticle has to have the opposite conserved property so that the property will remain conserved. As an example, the positron has the opposite charge as the electron so that electric charge is neither created or destroyed when electron-positron pairs appear or disappear.

While all particles have antiparticles, some particles like the photon, have no conserved properties other than energy. As a result, these particles are indistinguishable from their antiparticles.

#### PARTICLE NATURE OF FORCES

De Broglie got his idea for the wave nature of the electron from the particle-wave nature of light. The particle of light is the photon which can knock electrons out of a metal surface. The wave nature is the wave of electric and magnetic force that was predicted by Maxwell's theory. When you combine these two aspects of light, you are led to the conclusion that electric and magnetic forces are ultimately caused by photons. We call any force resulting from electric or magnetic forces as being due to the *electric interaction*. The photon is the particle responsible for the electric interaction.

Let us see how our picture of the hydrogen atom has evolved as we have learned more about the particles and forces involved. We started with a miniature solar system with the heavy proton at the center and an electron in orbit. The force was the electric force that in many ways resembled the gravitational force that keeps the earth in orbit around the sun. This picture failed, however, when we tried to explain the light radiated by heated hydrogen.

The next real improvement comes with Schrödinger's wave equation describing the behavior of the electron in hydrogen. Rather than there being allowed orbits as in Bohr's model, the electron in Schrödinger's picture has allowed *standing wave patterns*. The chemical properties of atoms can be deduced from these wave patterns, and Schrödinger's equation leads to accurate predictions of the wavelengths of light radiated not only by hydrogen but other atoms as well.

There are two limitations to Schrödinger's equation. One of the limitations we have seen is that it is a non relativistic equation, an equation that neglects any change in the electron's mass due to motion. While this is a very good approximation for describing the slow speed electron in hydrogen, the wavelengths of light radiated by hydrogen can be measured so accurately that tiny relativistic effects can be seen. Dirac's relativistic wave equation is required to explain these tiny relativistic corrections. The second limitation is that neither Schrödinger's or Dirac's equations take into account the particle nature of the electric force holding hydrogen together. In the hydrogen atom, the particle nature of the electric force has only the very tiniest effect on the wavelength of the radiated light. But even these effects can be measured and the particle nature must be taken into account. The theory that takes into account both the wave nature of the electron and the particle nature of the electric force is called *quantum electrodynamics*, a theory finally developed in 1947 by Richard Feynman and Julian Schwinger. Quantum electrodynamics is the most precisely tested theory in all of science.

In our current picture of the hydrogen atom, as described by quantum electrodynamics, the force between the electron and the proton nucleus is caused by the continual exchange of photons between the two charged particles. While being exchanged, the photon can do some subtle things like create a positron electron pair which quickly annihilates. These subtle things have tiny but measurable effects on the radiated wavelengths, effects that correctly predicted by the theory.

The development of quantum electrodynamics came nearly 20 years after Dirac's equation because of certain mathematical problems the theory had to overcome. In this theory, the electron is treated as a point particle with *no size*. The accuracy of the predictions of quantum electrodynamics is our best evidence that this is the correct picture. In other words, we have no evidence that the electron has a finite size, and a very accurate theory which assumes that it does not. However, it is not easy to construct a mathematical theory in which a finite amount of mass and energy is crammed into a region of no size. For one thing you are looking at infinite densities of mass and energy.

#### Renormalization

The early attempts to construct the theory of quantum electrodynamics were plagued by infinities. What would happen is that you would do an initial approximate calculation and the results would be good. You would then try to improve the results by calculating what were supposed to be tiny corrections, and the corrections turned out to be infinitely large. One of the main accomplishments of Feynman and Schwinger was to develop a mathematical procedure, sort of a mathematical slight of hand, that got rid of the infinities. This mathematical procedure became known as *renormalization*.

Feynman always felt that renormalization was simply a trick to cover up our ignorance of a deeper more accurate picture of the electron. I can still hear him saying this during several seminars. It turned out however that renormalization became an important guide in developing theories of other forces. We will shortly encounter two new forces as we look down into the atomic nucleus, forces called the *nuclear interaction* and the *weak interaction*. Both of these forces have a particle-wave nature like the electric interaction, and the successful theories of these forces used renormalization as a guide.



**Figure 8-33** Einstein's theory of gravity predicted that Mercury's elliptical orbit "precessed" or rotated somewhat like the rotation seen in the above orbit. Mercury's precession is much, much smaller.

#### Gravity

The one holdout, the one force for which we do not have a successful theory, is gravity. We have come a long way since Newton's law of gravity. After Einstein developed his theory of relativity in 1905, he spent the next 12 years working on a relativistic theory of gravity. The result, known as *general relativity* is a theory of gravity that is in many ways similar to Maxwell's theory of electricity. Einstein's theory predicts, for example, that a planet in orbit about a star should emit gravitational waves in much the same way that Maxwell's theory predicts that an electron in orbit about a nucleus should emit electromagnetic radiation or light.

One of the difficulties working with Einstein's theory of gravity is that Newton's theory of gravity explains almost everything we see, and you have to look very hard in places where Newton's law is wrong and Einstein's theory is right. There is an extremely small but measurable correction to the orbit of Mercury that Newton's theory cannot explain and Einstein's theory does.

Einstein's theory also correctly predicts how much light will be deflected by the gravitational attraction of a star. You can argue that because light has energy and energy is equivalent to mass, Newton's law of gravity should also predict that starlight should be deflected by the gravitational pull of a star. But this Newtonian argument leads to half the deflection predicted by Einstein's theory, and the deflection predicted by Einstein is observed.

The gravitational radiation predicted by Einstein's theory has not been detected directly, but we have very good evidence for its existence. In 1974 Joe Taylor from the University of Massachusetts, working at the large radio telescope at Arecibo discovered a pair of neutron stars in close orbit about each other. We will have more to say about neutron stars later. The point is that the period of the orbit of these stars can be measured with extreme precision.

Einstein's theory predicts that the orbiting stars should radiate gravitational waves and spiral in toward each other. This is reminiscent of what we got by applying Maxwell's theory to the electron in hydrogen, but in the case of the pair of neutron stars the theory worked. The period of the orbit of these stars is changing in exactly the way one would expect if the stars were radiating gravitational waves.

If our wave-particle picture of the behavior of matter is correct, then the gravitational waves must have a particle nature like electromagnetic waves. Physicists call the gravitational particle the *graviton*. We think we know a lot about the graviton even though we have not yet seen one. The graviton should, like the photon, have no rest mass, travel at the speed of light, and have the same relationship between energy and wavelength.

One difference is that because the graviton has energy and therefore mass, and because gravitons interact with mass, gravitons interact with themselves. This self interaction significantly complicates the theory of gravity. In contrast photons interact with electric charge, but photons themselves do not carry charge. As a result, photons do not interact with each other which considerably simplifies the theory of the electric interaction.

An important difference between the graviton and the photon, what has prevented the graviton from being detected, is its fantastically weak interaction with matter. You saw that the gravitational force between the electron and a proton is a thousand billion billion billion billion times weaker than the electric force. In effect this makes the graviton a thousand billion billion billion billion times harder to detect. The only reason we know that this very weak force exists at all is that it gets stronger and stronger as we put more and more mass together, to form large objects like planets and stars. Not only do we have problems thinking of a way to detect gravitons, we have run into a surprising amount of difficulty constructing a theory of gravitons. The theory would be known as the *quantum theory of gravity*, but we do not yet have a quantum theory of gravity. The problem is that the theory of gravitons interacting with point particles, the gravitational analogy of quantum electrodynamics, does not work. The theory is not renormalizable, you cannot get rid of the infinities. As in the case of the electric interaction the simple calculations work well, and that is why we think we know a lot about the graviton. But when you try to make what should be tiny relativistic corrections, the correction turns out to be infinite. No mathematical slight of hand has gotten rid of the infinities.

The failure to construct a consistent quantum theory of gravity interacting with point particles has suggested to some theoretical physicists that our picture of the electron and some other particles being point particles is wrong. In a new approach called *string theory*, the elementary particles are view not as point particles but instead as incredibly small one dimensional objects called *strings*. The strings vibrate, with different modes of vibration corresponding to different elementary particles.

String theory is complex. For example, the strings exist in a world of 10 dimensions, whereas we live in a world of 4 dimensions. To make string theory work, you have to explain what happened to the other six dimensions.

Another problem with string theory is that it has not led to any predictions that distinguish it from other theories. There are as yet no tests, like the deflection of starlight by the sun, to demonstrate that string theory is right and other theories are wrong.

String theory does, however, have one thing going for it. By spreading the elementary particles out from zero dimensions (points) to one dimensional objects (strings), the infinities in the theory of gravity can be avoided.

#### A SUMMARY

Up to this point our focus has been on the hydrogen atom. The physical magnification has not been too great, we are still picturing the atom as an object magnified to the size of a basketball with two particles, the electron and proton, that are too small to see. They may or may not have some size, but we cannot tell at this scale.

What we have done is change our perception of the atom. We started with a picture that Newton would recognize, of a small solar system with the massive proton at the center and the lighter electron held in orbit by the electric force. When we modernize the picture by including Maxwell's theory of electricity and magnetism, we run into trouble. We end up predicting that the electron will lose energy by radiating light, soon crashing into the proton. Bohr salvaged the picture by introducing his allowed orbits and quantized angular momentum, but the success of Bohr's theory only strengthened the conviction that something was fundamentally wrong with classical physics.

Louis de Broglie pointed the way to a new picture of the behavior of matter by proposing that all matter, not just light, had a particle-wave nature. Building on de Broglie's idea, Schrödinger developed a wave equation that not only describes the behavior of the electron in hydrogen, but in larger and more complex atoms as well.

While Schrödinger's non relativistic wave equation adequately explains most classical phenomena, even in the hydrogen atom, there are tiny but observable relativistic effects that Dirac could explain with his relativistic wave equation for the electron. Dirac handled the problem of all relativistic wave equations having two solutions by reinterpreting the second solution as representing antimatter.

Int-17

Dirac's equation is still not the final theory for hydrogen because it does not take into account the fact that electric forces are ultimately caused by photons. The wave theory of the electron that takes the photon nature of the electric force into account is known as quantum electrodynamics. The predictions of quantum electrodynamics are in complete agreement with experiment, it is the most precisely tested theory in science.

The problems resulting from treating the electron as a point particle were handled in quantum electrodynamics by renormalization. Renormalization does not work, however, when one tries to formulate a quantum theory of gravity where the gravitational force particle—the graviton—interacts with point particles. This has led some theorists to picture the electron not as a point but as an incredibly small one dimensional object called a *string*. While string theory is renormalizable, there have been no experimental tests to show that string theory is right and the point particle picture is wrong. This is as far as we can take our picture of the hydrogen atom without taking a closer look at the nucleus.



#### Figure 19-2 Isotopes of hydrogen and helium.

#### THE NUCLEUS

To see the nucleus we have to magnify our hydrogen atom to a size much larger than a basketball. When the atom is enlarged so that it would just fill a football stadium, the nucleus, the single proton, would be about the size of a pencil eraser. The proton is clearly not a point particle like the electron. If we enlarge the atom further to get a better view of the nucleus, to the point where the proton looks as big as a grapefruit, the atom is about 10 kilometers in diameter. This grapefruit sized object weighs 1836 times as much as the electron, but it is the electron wave that occupies the 10 kilometer sphere of space surrounding the proton.

Before we look inside the proton, let us take a brief look at the nuclei of some other atoms. Once in a great while you will find a hydrogen nucleus with two particles. One is a proton and the other is the electrically neutral particles called the *neutron*. Aside from the electric charge, the proton and neutron look very similar. They are about the same size and about the same mass. The neutron is a fraction of a percent heavier than the proton, a small mass difference that will turn out to have some interesting consequences.

As we mentioned, the type of element is determined by the number of protons in the nucleus. All hydrogen atoms have one proton, all helium atoms 2 protons, etc. But for the same element there can be different numbers of neutrons in the nucleus. Atoms with the same numbers of protons but different numbers of neutrons are called different *isotopes* of the element. Another isotope of hydrogen, one that is unstable and decays in roughly 10 years, is a nucleus with one proton and two neutrons called *tritium*.

The most stable isotope of helium is helium 4, with 2 protons and 2 neutrons. Helium 3 with 2 protons and one neutron is stable but very rare. Once we get beyond hydrogen we name the different isotopes by adding a number after the name, a number representing the total number of protons and neutrons. For example the heaviest, naturally occurring atom is the isotope Uranium 238, which has 92 protons and 146 neutrons for a total of 238 nuclear particles, or *nucleons* as we sometimes refer to them.

The nucleons in a nucleus pack together much like the grapes in a bunch, or like a bag of grapefruit. At our enlargement where a proton looks as big as a grapefruit, the uranium nucleus would be just over half a meter in diameter, just big enough to hold 238 grapefruit.

When you look at a uranium nucleus with its 92 positively charged protons mixed in with electrically neutral neutrons, then you have to wonder, what holds the thing together? The protons, being all positively charged, all repel each other. And because they are so close together in the nucleus, the repulsion is extremely strong. It is much stronger than the attractive force felt by the distant negative electrons. There must be another kind of force, and attractive force, that keeps the protons from flying apart.

The attractive force is not gravity. Gravity is so weak that it is virtually undetectable on an atomic scale. The attractive force that overpowers the electric repulsion is called the *nuclear force*. The nuclear force between nucleons is attractive, and essentially blind to the difference between a proton and a neutron. To the nuclear force, a proton and a neutron look the same. The nuclear force has no effect whatsoever on an electron.



**Figure 19-1** Styrofoam ball model of the uranium nucleus..

One of the important features of the nuclear force between nucleons is that it has a short range. Compared to the longer range electric force, the nuclear force is more like a contact cement. When two protons are next to each other, the attractive nuclear force is stronger than the electric repulsion. But separate the protons by more than about 4 protons diameters and the electric force is stronger.

If you make nuclei by adding nucleons to a small nucleus, the object becomes more and more stable because all the nucleons are attracting each other. But when you get to nuclei whose diameter exceeds around 4 proton diameters, protons on opposite sides of the nucleus start to repel each other. As a result nuclei larger than that become less stable as you make them bigger. The isotope Iron 56 with 26 protons and 30 neutrons, is about 4 proton diameters across and is the most stable of all nuclei. When you reach Uranium which is about 6 proton diameters across, the nucleus has become so unstable that if you jostle it by hitting it with a proton, it will break apart into two roughly equal sized more stable nuclei. Once apart, the smaller nuclei repel each other electrically and fly apart releasing electric potential energy. This process is called nuclear *fission* and is the source of energy in an atomic bomb.

While energy is released when you break apart the large unstable nuclei, energy is also released when you add nucleons to build up the smaller, more stable nuclei. For example, if you start with four protons (four hydrogen nuclei), turn two of the protons into neutrons (we will see how to do this shortly) and put them together to form stable helium 4 nucleus, you get a considerable release of energy. You can easily figure out how much energy is released by noting that 4 protons have a mass that is about .7 percent greater than a helium nucleus. As a result when the protons combine to form helium, about .7 percent of their mass is converted to other forms of energy. Our sun is powered by this energy release as it "burns" hydrogen to form helium. This process is called nuclear fusion and is the source of the energy of the powerful hydrogen bombs.

#### **STELLAR EVOLUTION**

Our sun is about half way through burning up the hydrogen in its hot, inner core. When the hydrogen is exhausted in another 5 billion years, the sun will initially cool and start to collapse. But the collapse will release gravitational potential energy that makes the smaller sun even hotter than it was before running out of hydrogen. The hotter core will emit so much light that the pressure of the light will expand the surface of the sun out beyond the earth's orbit, and the sun will become what is known as a red giant star. Soon, over the astronomically short time of a few million years, the star will cool off becoming a dying, dark ember about the size of the earth. It will become what is known as a *black dwarf*.

If the sun had been more massive when the hydrogen ran out and the star started to collapse, then more gravitational potential energy would have been released. The core would have become hotter, hot enough to ignite the helium to form the heavier nucleus carbon. Higher temperatures are required to burn helium because the helium nuclei, with two protons, repel each other with four times the electric repulsion than hydrogen nuclei. As a result more thermal energy is required to slam the helium nuclei close enough for the attractive nuclear force to take over.

Once the helium is burned up, the star again starts to cool and contract, releasing more gravitational potential energy until it becomes hot enough to burn the carbon to form oxygen nuclei. This cycle keeps repeating, forming one element after another until we get to Iron 56. When you have an iron core and the star starts to collapse and gets hotter, the iron does not burn. You do not get a release of energy by making nuclei larger than iron. As a result the collapse continues resulting in a huge implosion.

Once the center collapses, a strong shock wave races out through the outer layers of the star, tearing the star apart. This is called a *supernova explosion*. It is in these supernova explosions with their extremely high temperatures that nuclei larger than iron are formed. All the elements inside of you that are down the periodic table from iron were created in a supernova. *Part of you has already been through a supernova explosion*. What is left behind of the core of the star depends on how massive the star was to begin with. If what remains of the core is 1.4 times as massive as our sun, then the gravitational force will be strong enough to cram the electrons into the nuclei, turning all the protons into neutrons, and leaving behind a ball of neutrons about 20kilometers in diameter. This is called a*neutron star*. A neutron star is essentially a gigantic nucleus held together by gravity instead of the nuclear force.

If you think that squeezing the mass of a star into a ball 20 kilometers in diameter is hard to picture (at this density all the people on the earth would fit into the volume of a raindrop), then consider what happens if the remaining core is about six times as massive as the sun. With such mass, the gravitational force is so strong that the neutrons are crushed and the star becomes smaller and smaller.

The matter in a neutron star is about as rigid as matter can get. The more rigid a substance is, the faster sound waves travel through the substance. For example, sound travels considerably faster through steel than air. The matter in a neutron star is so rigid, or shall we say so incompressible, that the speed of sound approaches the speed of light.



**Figure 4** 1987 supernova as seen by the Hubble telescope.

When gravity has crushed the neutrons in a neutron star, it has overcome the strongest resistance any known force can possibly resist. But, as the collapse continues, gravity keeps getting stronger. According to our current picture of the behavior of matter, a rather unclear picture in this case, the collapse continues until the star becomes a point with no size. Well before it reaches that end, gravity has become so strong that light can no longer escape, with the result that these objects are known as **black holes**.

We have a fuzzy picture of what lies at the center of a black hole because we do not have a quantum theory of gravity. Einstein's classical theory of gravity predicts that the star collapses to a point, but before that happens we should reach a state where the quantum effects of gravity are important. Perhaps string theory will give us a clue as to what is happening. We will not learn by looking because light cannot get out.

The formation of neutron stars and black holes emphasizes an important feature of gravity. On an atomic scale, gravity is the weakest of the forces we have discussed so far. The gravitational force between an electron and a proton is a thousand billion billion billion billion  $(10^{39})$  times weaker than the electric force. Yet because gravity is long range like the electric force, and has no cancellation, it ends up dominating all other forces, even crushing matter as we know it, out of existence.



#### Figure 5

Hubble telescope's first view of a lone neutron star in visible light. This star is no greater than 16.8 miles (28 kilometers) across.

#### **The Weak Interaction**

In addition to gravity, the electric interaction and the nuclear force, there is one more basic force or interaction in nature given the rather bland name the *weak interaction*. While considerably weaker than electric or nuclear forces, it is far far stronger than gravity on a nuclear scale.

A distinctive feature of the weak interaction is its very short range. A range so short that only with the construction of the large accelerators since 1970 has one been able to see the weak interaction behave more like the other forces. Until then, the weak interaction was known only by reactions it could cause, like allowing a proton to turn into a neutron or vice versa.

Because of the weak interaction, an individual neutron is not stable. Within an average time of about 10 minutes it decays into a proton and an electron. Sometimes neutrons within an unstable nucleus also decay into a proton and electron. This kind of nuclear decay was observed toward the end of the nineteenth century when knowledge of elementary particles was very limited, and the electrons that came out in these nuclear decays were identified as some kind of a ray called a *beta ray*. (There were alpha rays which turned out to be helium nuclei, beta rays which were electrons, and gamma rays which were photons.) Because the electrons emitted during a neutron decay were called beta rays, the process is still known as the *beta decay* process.

The electron is emitted when a neutron decays in order to conerve electric charge. When the neutral neutron decays into a positive proton, a negatively charged particle must also be emitted so that the total charge does not change. The lightest particle available to carry out the negative charge is the electron.

Early studies of the beta decay process indicated that while electric charge was conserved, energy was not. For example, the rest mass of a neutron is nearly 0.14 percent greater than the rest mass of a proton. This mass difference is about four times larger than the rest mass of the electron, thus there is more than enough
mass energy available to create the electron when the neutron decays. If energy is conserved, you would expect that the energy left over after the electron is created would appear as kinetic energy of the electron.

Careful studies of the beta decay process showed that sometimes the electron carried out the expected amount of energy and sometimes it did not. These studies were carried out in the 1920s, when not too much was known about nuclear reactions. There was a serious debate about whether energy was actually conserved on the small scale of the nucleus.

In 1929, Wolfgang Pauli proposed that energy was conserved, and that the apparenty missing energy was carried out by an elusive particle that had not yet been seen. This elusive particle, which became known as the *neutrino* or "little neutral one", had to have some rather peculiar properties. Aside from being electrically neutral, it had to have essentially no rest mass because in some reactions the electron was seen to carry out all the energy, leaving none to create a neutrino rest mass.

The most bizarre property f the neutrino was its undetectability. It had to pass through matter leaving no trace. It was hard to believe such a particle could exist, yet on the other hand, it was hard to believe energy was not conserved. The neutrino was finally detected thirty years later and we are now quite confident that energy is conserved on the nuclear scale.

The neutrino is elusive because it interacts with matter only through the weak interaction (and gravity). Photons interact via the strong electric interaction and are quickly stopped when they encounter the electric charges in matter. Neutrinos can pass through light years of lead before there is a good chance that they will be stopped. Only in the collapsing core of an exploding star or in the very early universe is matter dense enough to significantly absorb neutrinos. Because neutrinos have no rest mass, they, like photons, travel at the speed of light.

## Leptons

We now know that neutrinos are emitted in the beta decay process because of another conservation law, the *conservation of leptons*. The leptons are a family of light particles that include the electron and the neutrino. When an electron is created, an anti neutrino is also created so that the number of leptons does not change.

Actually there are three distinct conservation laws for leptons. The lepton family consists of six particles, the electron, two more particles with rest mass and three different kinds of neutrino. The other massive particles are the *muon* which is 207 times as massive as the electron, and the recently discovered *tau* particle which is 3490 times heavier. The three kinds of neutrino are the *electron type neutrino*, the *muon type neutrino* and the *tau type neutrino*. The names come from the fact that each type of particle is separately conserved. For example when a neutron decays into a proton and an electron is created, it is an *anti electron type neutrino* that is created at the same time to conserve electron type particles.

In the other common beta decay process, where a proton turns into a neutron, a positron is created to conserve electric charge. Since the positron is the anti particle of the electron, its opposite, the electron type neutrino, must be created to conserve leptons.

# **Nuclear Structure**

The light nuclei, like helium, carbon, oxygen, generally have about equal numbers of protons and neutrons. As the nuclei become larger we find a growing excess of neutrons over protons. For example when we get up to Uranium 238, the excess has grown to 146 neutrons to 92 protons.

The most stable isotope of a given element is the one with the lowest possible energy. Because the weak interaction allows protons to change into neutrons and vice versa, the number of protons and neutrons in a nucleus can shift until the lowest energy combination is reached.

Two forms of energy that play an important role in their proess are the extra mass energy of the neutrons, and the electric potential energy of the protons. It takes a lot f to shove two protons together against their electric repulsion. The work you do in shoving them together is stored as electric potential energy which will be released if you let go and the particles fly apart. This energy will not be released, however, if the protons are latched together by the nuclear force. But in that case the electric potential energy can be released by turning one of the protons into a neutron. This will happen if enough electric potential energy is available not only to create the extra neutron rest mass energy, but also the positron required to conserve electric charge.

The reason that the large nuclei have an excess of neutrons over protons is that electric potential energy increases faster with increasing number of protons than neutron mass energy does with increasing numbers of neutrons. The amount of extra neutron rest mass energy is more or less proportional to the number of neutrons. But the increase in electric potential energy as you add a proton depends on the number of protons already in the nucleus. The more protons already there, the stronger the electric repulsion when you try to add another proton, and the greater the potential energy stored. As a result of this increasing energy cost of adding more protons, the large nuclei find their lowest energy balance having an excess of neutrons.

# **A CONFUSING PICTURE**

By 1932, the basic picture of matter looked about as simple as it can possibly get. The elementary particles were the proton, neutron, and electron. Protons and neutrons were held together in the nucleus by the nuclear force, electrons were bound to nuclei by the electric force to form atoms, a residual of the electric force held atoms together to form molecules, crystals and living matter, and gravity held large chunks of matter together for form planets, stars and galaxies. The rules governing the behavior of all this was quantum mechanics on a small scale, which became Newtonian mechanics on the larger scale of our familiar world. There were a few things still to be straightened out, such as the question as to whether energy was conserved in beta decays, and in fact why beta decays occurred at all, but it looked as if these loose ends should be soon tied up.

The opposite happened. By 1960, there were well over 100 so called elementary particles, all of them unstable except for the familiar electron, proton and neutron. Some lived long enough to travel kilometers down through the earth's atmosphere, others long enough to be observed in particle detectors. Still others had such short lifetimes that, even moving at nearly the speed of light, they could travel only a few proton diameters before decaying. With few exceptions, these particles were unexpected and their behavior difficult to explain. Where they were expected, they were incorrectly identified.

One place to begin the story of the progression of unexpected particles is with a prediction made in 1933 by Heidi Yukawa. Yukawa proposed a new theory of the nuclear force. Noting that the electric force was ultimately caused by a particle, Yukawa proposed that the nuclear force holding the protons and neutrons together in the nucleus was also caused by a particle, a particle that became known as the nuclear force meson. The zero rest mass photon gives rise to the long range electric force. Yukawa developed a wave equation for the nuclear force meson in which the range of the force depends on the rest mass of the meson. The bigger the rest mass of the meson, the shorter the range. (Later in the text, we will use the uncertainty principle to explain this relationship between the range of a force and the rest mass of the particle causing it.)

From the fact that iron is the most stable nucleus, Yukawa could estimate that the range of the nuclear force is about equal to the diameter of an iron nucleus, about four proton diameters. From this, he predicted that the nuclear force meson should have a rest mass bout 300 times the rest mass of the electron (about 1/6 the rest mass of a proton).

Shortly after Yukawa's prediction, the muon was discovered in the rain of particles that continually strike the earth called cosmic rays. The rest mass of the muon was found to be about 200 times that of the electron, not too far off the predicted mass of Yukawa's particle. For a while the muon was hailed as Yukawa's nuclear force meson. But further studies showed that muons could travel considerable distances through solid matter. If the muon were the nuclear force meson, it should interact strongly with nuclei and be stopped rapidly. Thus the muon was seen as not being Yukawa's particle. Then there was the question of what role the muon played. Why did nature need it?

In 1947 another particle called the  $\pi$  meson was discovered. (There were actually three  $\pi$  mesons, one with a positive charge, the  $\pi^+$ , one neutral, the  $\pi^\circ$ , and one with a negative charge, the  $\pi^-$ .) The  $\pi$  mesons interacted strongly with nuclei, and had the mass close to that predicted by Yukawa, 274 electron masses. The  $\pi$  mesons were then hailed as Yukawa's nuclear force meson.

However, at almost the same time, another particle called the K meson, 3.5 times heavier than the  $\pi$  meson, was discovered. It also interacted strongly with nuclei and clearly played a role in the nuclear force. The nuclear force was becoming more complex than Yukawa had expected.

Experiments designed to study the  $\pi$  and K mesons revealed other particles more massive than protons and neutrons that eventually decayed into protons and neutrons. It became clear that the proton and neutron were just the lightest members of a family of proton like particles. The number of particles in the proton family was approaching 100 by 1960. During this time it was also found that the  $\pi$  and K mesons were just the lightest members of another family of particles whose number exceeded 100 by 1960. It was rather mind boggling to think of the nuclear force as being caused by over 100 different kinds of mesons, while the electric force had only one particle, the photon.

One of the helpful ways of viewing matter at that time was to identify each of the particle decays with one of the four basic forces. The very fastest decays were assumed to be caused by the strong nuclear force. Decays that were about 100 times slower were identified with the slightly weaker electric force. Decays that took as long as a billionth of a second, a relatively long lifetime, were found to be caused by the weak interaction. The general scheme was the weaker the force, the longer it took to cause a particle decay.



#### Figure 6

First bubble chamber photograph of the  $\Omega^$ particle. The  $\Omega^-, \Xi^0, \Lambda^0$ and p<sup>+</sup> are all members of the proton family, the K's and  $\pi$ 's are mesons, the  $\gamma$ 's are photons and the e<sup>-</sup> and e<sup>+</sup> are electrons and positrons. Here we see two examples of the creation of an electron-positron pair by a photon.



# QUARKS

The mess seen in 1960 was cleaned up, brought into focus, primarily by the work of Murray Gell-Mann. In 1961 Gell-Mann and Yval Neuman found a scheme that allowed one to see symmetric patterns in the masses and charges of the various particles. In 1964 Gell-Mann and George Zweig discovered what they thought was the reason for the symmetries. The symmetries would be the natural result if the proton and meson families of particles were made up of smaller particles which Gell-Mann called *quarks*.

Initially Gell-Mann proposed that there were three different kinds of quark, but the number has since grown to six. The lightest pair of the quarks, the so called *up quark* and *down quark* are found in protons and neutrons. If the names "up quark" and "down quark" seem a bit peculiar, they are not nearly as confusing as the names *strange quark*, *charm quark*, *bottom quark* and *top quark* given the other four members of the quark family. It is too bad that the Greek letters had been used up naming other particles.

In the quark model, all members of the proton family consist of three quarks. The proton and neutron, are made from the *up* and *down* quarks. The proton consists of two *up* and one *down* quark, while the neutron is made from one *up* and two *down* quarks. The weak interaction, which as we saw can change protons into neutrons, does so by changing one of the proton's *up* quarks into a *down* quark.

The  $\pi$  meson type of particles, which were thought to be Yukawa's nuclear force particles, turned out instead to be quark-antiquark pairs. The profusion of what were thought to be elementary particles in 1960 resulted from the fact that there are many ways to combine three quarks to produce members of the proton family or a quark and an antiquark to create a meson. The fast elementary particle reactions were the result of the rearrangement of the quarks within the particle, while the slow reactions resulted when the weak interaction changed one kind of quark into another.

A peculiar feature of the quark model is that quarks have a *fractional charge*. In all studies of all elementary particles, charge was observed to come in units of the amount of charge on the electron. The electron had (-1) units, and the neutron (0) units. All of the more than 100 "elementary" particles had either +1, 0, or -1 units of change. Yet in the quark model, quarks had a charge of either (+2/3) units like the *up* quark or (-1/3) units like the *down* quark. (The anti particles have the opposite charge, -2/3 and +1/3 units respectively.) You can see that a proton with two *up* and one *down* quark has a total charge of (+2/3 +2/3 -1/3) = (+1) units, and the neutron with two *down* and one *up* quark has a total charge (-1/3 -1/3 +2/3) = (0) units.

The fact that no one had ever detected an individual quark, or ever seen a particle with a fractional charge, made the quark model hard to accept at first. When Gell-Mann initially proposed the model in 1963, he presented it as a mathematical construct to explain the symmetries he had earlier observed.

The quark model gained acceptance in the early 1970s when electrons at the Stanford high energy accelerator were used to probe the structure of the proton. This machine had enough energy, could look in sufficient detail to detect the three quarks inside. The quarks were real.

In 1995, the last and heaviest of the six quarks, the *top* quark, was finally detected at the Fermi Lab Accelerator. The *top* quark was difficult to detect because it is 185 times as massive as a proton. A very high energy accelerator was needed to create and observe this massive particle.

With the quark model, our view of matter has become relatively simple again: there are two families of particles called quarks and leptons. Each family contains six particles. It is not a coincidence that there are the same number of particles in each family. In the current theory of matter called the *standard model*, each pair of leptons is intimately connected to a pair of quarks. The electron type leptons are associated with the *down* and *up* quarks, the muon and muon type neutrino with the *strange* and *charm* quarks, and the tau and tau type neutrino with the *bottom* and *top* quarks.

Are there more than six quarks and six leptons? Are there still heavier lepton neutrino pairs associated with still heavier quarks? That the answer is no, that six is the limit, first came not from accelerator experiments, but from studies of the early universe. Here we have a question concerning the behavior of matter on the very smallest of scales of distance, at the level of quarks inside proton like particles, and we find the answer by looking at matter on the very largest of scales, the entire universe. The existence of more than six leptons and quarks would have altered the relative abundance of hydrogen, deuterium, and helium remaining after the big bang. It would have led to an abundance that is not consistent with what we see now. Later experiments with particle accelerators confirmed the results we first learned from the early universe.

Our picture of the four basic interactions has also become clearer since the early 1930s. The biggest change is in our view of the nuclear force. The basic nuclear force is now seen to be the force between quarks that holds them together to form protons, neutrons and other particles. What we used to call the nuclear force, that short range force binding protons and neutrons together in a nucleus, is now seen as a residual effect of the force between quarks. The old nuclear force is analogous to the residual electric force that binds complete atoms together to form molecules.

As the electric interaction is caused by a particle, the photon, the nuclear force is also caused by particles, eight different ones called *gluons*. The nuclear force is much more complex than the electric force because gluons not only interact with quarks, they also interact with themselves. This gives rise to a very strange force between quarks. Other forces get weaker as you separate the interacting particles. The nuclear force between quarks gets stronger! As a result quarks are confined to live inside particles like protons, neutrons and mesons. This is why we have still never seen an individual quark or an isolated particle with a fractional charge.



**Figure 28-29** Fermi Lab accelerator where the top quark was first observed.



**Figure 28-28** *Fermi Lab accelerator magnets.* 

# THE ELECTROWEAK THEORY

Another major advance in our understanding of the nature of the basic interactions came in 1964 when Steven Weinberg, Abdus Salam and Sheldon Glashow discovered a basic connection between the electric and weak interactions. Einstein had spent the latter part of his life trying without success to unify, find a common basis for, the electric and the gravitational force. It came somewhat as a surprise that the electric and weak interactions, which appear so different, had common origins. Their theory of the two forces is known as the *electroweak theory*.

In the electroweak theory, if we heat matter to a temperature higher than 1000 billion degrees, we will find that the electric and weak interaction are a single force. If we then let the matter cool, this single electroweak force splits into the two separate forces, the electric interaction and the weak interaction. This splitting of the forces is viewed as a so called *phase transition*, a transition in the state of matter like the one we see when water turns to ice at a temperature of 0°C. The temperature of the phase transition for the electroweak force sounds impossibly hot, but it is attainable if we build a big enough accelerator. The

cancelled superconducting supercollider was supposed to allow us to study the behavior of matter at these temperatures.

One of the major predictions of the electroweak theory was that after the electric and weak interactions had separated, electric forces should be caused by zero rest mass photons and the weak interaction should be caused by three rather massive particles given the names  $W^+$ ,  $W^-$  and  $Z^0$  mesons. These mesons were found, at their predicted mass, in a series of experiments performed at CERN in the late 1970s.

We have discussed Yukawa's meson theory of forces, a theory in which the range of the force is related to the rest mass of the particle responsible for the force. As it turns out, Yukawa's theory does not work for nuclear forces for which it was designed. The gluons have zero rest mass but because of their interaction, gives rise to a force unlike any other. What Yukawa's theory does describe fairly well is the weak interaction. The very short range of the weak interaction is a consequence of the large masses of the weak interaction mesons W<sup>+</sup>, W<sup>-</sup> and Z<sup>0</sup>. (The W mesons are 10 times as massive as a proton, the Z<sup>0</sup> is 11 times as massive.)



**Figure 28-30** Paths for the large particle accelerators at CERN. The Geneva airport is in the foreground.

# THE EARLY UNIVERSE

In the reverse motion picture of the expanding universe, the universe becomes smaller and smaller and hotter as we approach the big bang that created it. How small and how hot are questions we are still studying. But it now seems that with reasonable confidence we can apply the laws of physics to a universe that is about one nanosecond old and at a temperature of three hundred thousand billion degrees. This is the temperature of the electroweak transition where the weak and electric interactions become separate distinct forces. We have some confidence in our knowledge of the behavior of matter at this temperature because this temperature is being approached in the largest of the particle accelerators.

## $3 \times 10^{14}$ degrees

At three hundred thousand billion degrees the only structures that survive the energetic thermal collisions are the elementary particles themselves. At this time the universe consists of a soup of quarks and anti quarks, leptons and anti leptons, gravitons and gluons. Photons and the weak interaction mesons  $W^+$ ,  $W^-$  and  $Z^\circ$  are just emerging from the particle that gave rise to the electroweak force. The situation may not actually be that simple. When we get to that temperature we may find some of the exotic elementary particles suggested by some recent attempts at a quantum theory of gravity.

# 10<sup>13</sup> degrees

When the universe reaches the ripe old age of a millionth of a second, the time it takes light to travel 1000 feet, the temperature has dropped to 10 thousand billion degrees. At these temperatures the gluons are able to hold the quarks together to form protons, neutrons, mesons, and their anti particles. It is still much too hot, however, for protons and neutrons to stick together to form nuclei.

When we look closely at the soup of particles at 10 thousand billion degrees, there is activity in the form of the annihilation and creation of particle-antiparticle pairs. Proton-antiproton pairs, for example, are rapidly annihilating, turning into photons and mesons. But just as rapidly photons and mesons are creating proton-antiproton pairs.

In the next 10 millionths of a second the universe expands and cools to a point where the photons and mesons no longer have enough energy to recreate the rapidly annihilating proton and neutron pairs. Soon the protons and neutrons and their antiparticles will have essentially disappeared from the universe.

#### **Matter particles survive**

The protons and neutrons will have almost disappeared but not quite. For some reason, not yet completely understood, there was a tiny excess of protons over antiprotons and neutrons over antineutrons. The estimate is that there were 100,000,000,001 matter particles for every 100,000,000,000 antimatter particles. It was the tiny excess of matter over antimatter that survived the proton and neutron annihilation.

## 3×10<sup>10</sup> degrees

After this annihilation, nothing much happens until the universe approaches the age of a tenth of a second and the temperature has dropped to 30 billion degrees. During this time the particles we see are photons, neutrinos and antineutrinos and electrons and positrons. These particles exist in roughly equal numbers. The electron-positron pairs are rapidly annihilating to produce photons, but the photons are equally rapidly creating electron positron pairs.

#### 38% neutrons

There are still the relatively few protons and neutrons that survived the earlier annihilation. The weak interaction allows the protons to turn into neutrons and vice versa, with the result there are roughly equal numbers of protons and neutrons. The numbers are not quite equal, however, because at those temperatures there is a slightly greater chance for the heavier neutron to decay into a lighter proton than vice versa. It is estimated that the ratio of neutrons to protons has dropped to 38% by the time the universe is .11 seconds old. The temperature is still too high for protons and neutrons to combine to form nuclei.

#### Neutrinos escape at one second

As we noted, neutrinos are special particles in that the only way they interact with matter is through the weak interaction. Neutrinos pass right through the earth with only the slightest chance of being stopped. But the early universe is so dense that the neutrinos interact readily with all the other particles.

When the universe reaches an age of about one second, the expansion has reduced the density of matter to the point that neutrinos can pass undisturbed through matter. We can think of the neutrinos as decoupling from matter and going on their own independent way. From a time of one second on, the only thing that will happen to the neutrinos is that they will continue to cool as the universe expands. At an age of 1 second, the neutrinos were at a temperature of 10 billion degrees. By today they have cooled to only a few degrees above absolute zero. This is our prediction, but these cool neutrinos are too elusive to have been directly observed.

## 24% neutrons

Some other interesting things are also beginning to happen at the time of 1 second. The photons have cooled to a point that they just barely have enough energy to create electron-positron pairs to replace those that are rapidly annihilating. The result is that the electrons and positrons are beginning to disappear. At these temperatures it is also more favorable for neutrons to turn into protons rather than vice versa, with the result that the ratio of neutrons to protons has dropped to 24%.

# 3×10<sup>9</sup> degrees (13.8 seconds)

When the temperature of the universe has dropped to 3 billion degrees, at the time of 13.8 seconds, the energy of the photons has dropped below the threshold of being able to create electron-positron pairs and the electrons and the positrons begin to vanish from the universe. There was the same tiny excess of electrons over positrons as there had been of protons over antiprotons. Only the excess of electrons will survive.

## **Positrons annihilated**

After about three minutes the positrons are gone and from then on the universe consists of photons, neutrinos, anti neutrinos and the few matter particles. The neutrinos are not interacting with anything, and the matter particles are outnumbered by photons in a ration of 100,000,000,000 to one. The photons essentially dominate the universe.

## **Deuterium bottleneck**

At the time of 13.8 seconds the temperature was 3 billion degrees, cool enough for helium nuclei to survive. But helium nuclei cannot be made without first making deuterium, and deuterium is not stable at that temperature. Thus while there are still neutrons around, protons and neutrons still cannot form nuclei because of this deuterium bottleneck.

# Helium created

When the universe reaches an age of three minutes and 2 seconds, and the ratio of neutrons to protons has dropped to 13%, finally deuterium is stable. These surviving neutrons are quickly swallowed up to form deuterium which in turn combine to form the very stable helium nuclei. Since there are equal numbers of protons and neutrons in a helium nucleus, the 13% of neutrons combined with an equal number of protons to give 26% by weight of helium nuclei and 74% protons or hydrogen nuclei.

By the time the helium nuclei form, the universe has become too cool to burn the helium to form heavier elements. The creation of the heavier elements has to wait until stars begin to form one third of a million years later.

The formation of elements inside of stars was the basis of the continuous creation theory. As we mentioned, one could explain the abundance of all the elements except helium as being a by product of the evolution of stars. To explain the helium abundance it was necessary to abandon his continuous creation theory and accept that there might have been a big bang after all.

# **The Thermal Photons**

After the electron positron pairs had vanished, what is left in the universe are the photons, neutrinos, anti neutrinos, and the few matter particles consisting of protons, helium nuclei and a trace of deuterium and lithium. There are enough electrons to balance the charge on the hydrogen and helium nuclei, but the photons are energetic enough to break up any atoms that might try to form. The neutrinos have stopped interacting with anything and the matter particles are outnumbered by photons in a ratio of 100 billion to one. At this time the photons dominate the universe.

One way to understand why the universe cools as it expands is to picture the expansion of the universe as stretching the wavelength of the photons. Since the energy of a photon is related to its wavelength (the longer the wavelength the lower the energy), this stretching of wavelengths lowers the photon energies. Because the photons dominate the young universe, when the photons lose energy and cool down, so does everything else that the photons are interacting with.

#### .7 million years

Until the universe reaches the age of nearly a million years, the photons are knocking the matter particles around, preventing them from forming whole atoms or gravitational structures like stars. But at the age of .7 million years the temperature has dropped to 3000 degrees, and something very special happens at that point. The matter particles are mostly hydrogen, and if you cool hydrogen below 3000 degrees it becomes transparent. The transition in going from above 3000 degrees to below, is like going from inside the surface of the sun to outside. We go from an opaque, glowing universe to a transparent one.

#### **Transparent universe**

When the universe becomes transparent, the photons no longer have any effect on the matter particles and the matter can begin to form atoms, stars, and galaxies. Everything we see today, except for the primordial hydrogen and helium, was formed after the universe became transparent. Think about what it means that the universe became transparent at an age of .7 million years. In our telescopes, as we look at more and more distant galaxies, the light from these galaxies must have taken more and more time to reach us. As we look farther out we are looking farther back in time. With the Hubble telescope we are now looking at galaxies formed when the universe was less than a billion years old, less than 10% of its current age.

Imagine that you could build a telescope even more powerful than the Hubble, one that was able to see as far out, as far back to when the universe was .7 million years old. If you could look that far out what would you see? You would be staring into a wall of heated opaque hydrogen. You would see this wall in every direction you looked. If you tried to see through the wall, you would be trying to look at the universe at earlier, hotter times. It would be as futile as trying to look inside the sun with a telescope.

Although this wall at .7 million light years consists of essentially the same heated hydrogen as the surface of the sun, looking at it would not be the same as looking at the sun. The light from this wall has been traveling toward us for the last 14 billion years, during which time the expansion of the universe has stretched the wavelength and cooled the photons to a temperature of less than 3 degrees, to a temperature of 2.74 degrees above absolute zero to be precise.

Photons at a temperature of 2.74 degrees can be observed, not by optical telescopes but by radio antennas instead. In 1964 the engineers Arno Penzias and Robert Wilson were working with the radio antenna that was communicating with the Telstar satellite. The satellite was a large aluminized balloon that was supposed to reflect radio signals back to earth. The radio antenna had to be very sensitive to pick up the weak reflected signals. In checking out the antenna, Penzias and Wilson were troubled by a faint noise that they could not eliminate. Further study showed that the noise was characteristic of a thermal bath of photons whose temperature was around 3 degrees. After hearing a seminar on the theory of the big bang and on the possibility that there might be some light remaining from the explosion, Penzias and Wilson immediately realized that the noise in their antenna was that light. Their antenna in effect was looking at light from the time the universe became transparent. At that time, only a few astronomers and physicists were taking the big bang hypothesis seriously. The idea of the universe beginning in an explosion seemed too preposterous. After Penzias and Wilson saw the light left over from the hotter universe, no other view has been acceptable.

The fact that the universe became transparent at an age of .7 million years, means that the photons, now called the *cosmic background radiation*, travelled undisturbed by matter. By studying these photons carefully, which we are now doing in various rocket and satellite experiments, we are in a sense, taking an accurate photograph of the universe when the universe was .7 million years old.



**Figure 34-11** *Penzias and Wilson, and the Holmdel radio telescope.* 

This photograph shows an extremely uniform universe. The smoothness shows us that stars and galaxies had not yet begun to form. In fact the universe was so smooth that it is difficult to explain how galaxies did form in the time between when the universe went transparent and when we see galaxies in the most distant Hubble telescope photographs. The COBE (Cosmic Background Explorer) satellite was able to detect tiny fluctuations in the temperatures of the background radiation, indicating that there was perhaps just enough structure in the early hot universe to give us the stars, galaxies and clustering of galaxies we see today.

One of the questions you may have had reading our discussion of the early universe, is how do we know that the photons, and earlier the particle-anti particle pairs outnumbered the matter particles by a ratio of 100 billion to one? How do we estimate the tiny excess of matter over anti matter that left behind all the matter we see today? The answer is that the thermal photons we see today outnumber protons and neutrons by a factor of 100 billion to one and that ratio should not have changed since the universe was a few minutes old.

We also mentioned that it would be futile to try to look under the surface of the sun using a telescope. That is true if we try to use a photon telescope. However we can, in effect, see to the very core of the sun using neutrinos. In the burning of hydrogen to form helium, for each helium nucleus created, two protons are converted to neutrons via the weak interaction. In the process two neutrinos are emitted. As a result the core of the sun is a bright source of neutrinos which we can detect and study here on earth.

While it would be futile to use photons to see farther back to when the universe was about .7 million years old, we should be able to see through that barrier using neutrinos. The universe became transparent to neutrinos at the end of the first second. If we could detect these neutrinos, we would have a snapshot of the universe as it looked when it was one second old. Thus far, we have not found a way to detect these cosmic background neutrinos.

# Chapter 1 Principle of Relativity

The subject of this book is the behavior of matter—the particles that make up matter, the interactions between particles, and the structures that these interactions create. There is a wondrous variety of activity, as patterns and structures form and dissipate, and all of this activity takes place in an arena we call space and time. The subject of this chapter is that arena—space and time itself.

Initially, one might think that a chapter on space and time would either be extraordinarily dull, or too esoteric to be of any use. From the **it's too dull** point of view, distance is measured by meter sticks, and there are relationships like the Pythagorean theorem and various geometric and trigonometric rules already familiar to the reader. Time appears to be less challenging—it is measured by clocks and seems to march inexorably forward.

On the **too esoteric** side are the theories like Einstein's General Theory of Relativity which treats gravity as a distortion of space and time, the Feynman-Wheeler picture of antimatter as being matter traveling backward in time, and recent "super symmetry" theories which assume a ten dimensional space. All of these theories are interesting, and we will briefly discuss them. We will do that later in the text after we have built up enough of a background to understand why these theories were put forth.

What can we say in an introductory chapter about space and time that is interesting, or useful, or necessary for a physics text? Why not follow the traditional approach and begin with the development of Newton's theory of mechanics. You do not need a very sophisticated picture of space and time to understand Newtonian mechanics, and this theory explains an enormous range of phenomena, more than you can learn in one or several years. There are three main reasons why we will not start off with the Newtonian picture. The first is that the simple Newtonian view of space and time is approximate, and the approximation fails badly in many examples we will discuss in this text. By starting with a more accurate picture of space and time, we can view these examples as successful predictions rather than failures of the Newtonian theory.

The second reason is that the more accurate picture of space and time is based on the simplest, yet perhaps most general law in all of physics—the principle of relativity. The principle of relativity not only underlies all basic theories of physics, it was essential in the discovery of many of these theories. Of all possible ways matter could behave, only a very, very few are consistent with the principle of relativity, and by concentrating on these few we have been able to make enormous strides in understanding how matter interacts. By beginning the text with the principle of relativity, the reader starts off with one of the best examples of a fundamental physical law.

Our third reason for starting with the principle of relativity and the nature of space and time, is that it is fun. The math required is simple–only the Pythagorean theorem. Yet results like clocks running slow, lengths contracting, the existence of an ultimate speed, and questions of causality, are stimulating topics. Many of these results are counter intuitive. Your effort will not be in struggling with mathematical formulas, but in visualizing yourself in new and strange situations. This visualization starts off slowly, but you will get used to it and become quite good at it. By the end of the course the principle of relativity, and the consequences known as Einstein's special theory of relativity will be second nature to you.

# THE PRINCIPLE OF RELATIVITY

In this age of jet travel, the principle of relativity is not a strange concept. It says that *you cannot feel motion in a straight line at constant speed*. Recall a smooth flight where the jet you were in was traveling at perhaps 500 miles per hour. A moving picture is being shown and all the window shades are closed. As you watch the movie are you aware of the motion of the jet? Do you feel the jet hurtling through the air at 500 miles per hour? Does everything inside the jet crash to the rear of the plane because of this immense speed?

No-the only exciting thing going on is the movie. The smooth motion of the jet causes no excitement whatsoever. If you spill a diet Coke, it lands in your lap just as it would if the plane were sitting on the ground. The problem with walking around the plane is the food and drink cart blocking the aisles, not the motion of the plane. Because the window shades are closed, you cannot even be sure that the plane is moving. If you open your window shade and look out, and if it is daytime and clear, you can look down and see the land move by. Flying over the Midwestern United States you will see all those square 40 acre plots of land move by, and this tells you that you are moving. If someone suggested to you that maybe the farms were moving and you were at rest, you would know that was ridiculous, the plane has the jet engines, not the farms.

Despite the dull experience in a jumbo jet, we often are able to sense motion. There is no problem in feeling motion when we start, stop, or go around a sharp curve. But starting, stopping, and going around a curve are not examples of motion at constant speed in a straight line, the kind of motion we are talking about. *Changes in speed or in direction of motion are called accelerations*, and we can feel accelerations. (Note: In physics a decrease in speed is referred to as a negative acceleration.)

Even without accelerations, even when we are moving at constant speed in a straight line, we can have a strong sense of motion. Driving down a freeway at 60 miles per hour in a low-slung, open sports car can be a notable, if not scary, experience. This sense of motion can be misleading. The first wide screen moving pictures took the camera along on a roller coaster ride. Most people in the audience found watching this ride to be almost as nerve wracking as actually riding a roller coaster. Some even became sick. Yet the audience was just sitting at rest in the movie theater.

## Exercise 1

Throughout this text we will insert various exercises where we want you to stop and think about or work with the material. At this point we want you to stop reading and think about various times you have experienced motion. Then eliminate all those that involved accelerations, where you speeded up, slowed down, or went around a curve. What do you have left, and how real were the sensations?

One of my favorite examples occurred while I was at a bus station in Boston. A number of busses were lined up side by side waiting for their scheduled departure times. I recall that after a fairly long wait, I observed that we were moving past the bus next to us. I was glad that we were finally leaving. A few seconds later I looked out the window again; the bus next to us had left and we were still sitting in the station. I had mistaken that bus's motion for our own!

# A Thought Experiment

Not only can you feel accelerated motion, you can easily see relative motion. I had no problem seeing the bus next to us move relative to us. My only difficulty was in telling whether they were moving or we were moving.

An example of where it is more obvious who is moving is the example of the jet flying over the Midwestern plains. In the daytime the passengers can see the farms go by; it is easy to detect the relative motion of the plane and the farms. And it is quite obvious that it is the plane moving and the farms are at rest. Or is it?

To deal with this question we will go through what is called a **thought experiment** where we solve a problem by imagining a sometimes contrived situation, and then figure out what the consequences would be if we were actually in that situation. Galileo is well known for his use of thought experiments to explain the concepts of the new mechanics he was discovering. For our thought experiment, imagine that we are going to take the Concorde supersonic jet from Boston, Massachusetts to San Francisco, California. The jet has been given special permission to fly across the country at supersonic speeds so that the trip, which is scheduled to leave at noon, takes only three hours.

When we arrive in San Francisco we reset our watches to Pacific Standard Time to make up for the 3 hour difference between Boston and San Francisco. We reset our watches to noon. When we left, it was noon and the sun was overhead. When we arrive it is still noon and the sun is still overhead. One might say that the jet flew fast enough to follow the sun, the 3 hour trip just balancing the 3 hours time difference.

But there is another view of the trip shown in Figure (1). When we took off at noon, the earth, the airplane and sun were lined up as shown in Figure (1a). Three hours later the earth, airplane and sun are still lined up as shown in Figure (1b). The only difference between (1a) and (1b) is that the earth has been rotating for three hours so that San Francisco, rather than Boston is now under the plane. The view in Figure (1) is what an astronaut approaching the earth in a spacecraft might see.



three hours later

over San Francisco.

a) Supersonic jet over Boston just after takeoff.

#### Figure 1

One view of a three hour trip from Boston to San Francisco. It is possible, even logical, to think of the jet as hovering at rest while the earth turns underneath. For someone inside the jet, looking down at the Midwestern farms going by, who is really moving? Are the farms *really* at rest and the plane moving? Or is the plane at rest and the farms going by? Figure (1) suggests that the latter point of view may be more accurate, at least from the perspective of one who sees the bigger picture including the earth, airplane, and sun.

But, you might ask, what about the jet engines and all the fuel that is being expended to move the jet at 1000 miles/ hour? Doesn't that prove that it is the jet that is moving? Not necessarily. When the earth rotates, it drags the atmosphere around with it creating a 1000 mi/hr wind that the plane has to fly through in order to stand still. Without the jet engines and fuel, the plane would be dragged back with the land and never reach San Francisco.

This thought experiment has one purpose. To loosen what may have been a firmly held conviction that when you are in a plane or car, you are moving and the land that you see go by must necessarily be at rest. Perhaps, under some circumstances it is more logical to think of yourself at rest and the ground as moving. *Or, perhaps it does not make any difference*. The principle of relativity allows us to take this last point of view.

# Statement of the Principle of Relativity

Earlier we defined uniform motion as motion at constant speed in a straight line. And we mentioned that the principle of relativity said that you could not feel this uniform motion. Since it is not exactly clear what is meant by "feeling" uniform motion, a more precise statement of the principle of relativity is needed, a statement that can be tested by experiment. The following is the definition we will use in this text.

> Imagine that you are in a capsule and you may have any equipment you wish inside the capsule. The principle of relativity states that there is no experiment you can perform that will allow you to tell whether or not the capsule is moving with uniform motion—motion in a straight line at constant speed.

In the above definition the capsule can use anything you want as an example—a jet plane, a car, or a room in a building. Generally, think of it as a sealed capsule like the jet plane where the moving picture is being shown and all of the window shades are shut. Of course you can look outside, and you may see things going by. But, as shown in Figure (1), seeing things outside go by does not prove that you and the capsule are moving. That cannot be used as evidence of your own uniform motion.

Think about what kind of experiments you might perform in the sealed capsule to detect your uniform motion. One experiment is to drop a coin on the floor. If you are at rest, the coin falls straight down. But if you are in a jet travelling 500 miles per hours and the flight is smooth, and you drop a coin, the coin still falls straight down. Dropping a coin does not distinguish between being at rest or moving at 500 miles per hour; this is one experiment that does not violate the principle of relativity.

There are many other experiments you can perform. You could use gyroscopes, electronic circuits, nuclear reactions, gravitational wave detectors, anything you want. The principle of relativity states that none of these will allow you to detect your uniform motion.

# Exercise 2

Think about what you might put inside the capsule and what experiments you might perform to detect the motion of the capsule. Discuss your ideas with others and see if you can come up with some way of violating the principle of relativity.

# **Basic Law of Physics**

We mentioned that one of the incentives for beginning the text with the principle of relativity is that it is an excellent example of a basic law of physics. It is simple and easy to state—there is no experiment that you can perform that allows you to detect your own uniform motion. Yet it is general—there is no experiment that can be done at any time, at any place, using anything, that can detect your uniform motion. And most important, it is completely subject to experimental test on an all-or-nothing basis. Just one verifiable experiment detecting one's own uniform motion, and the principle of relativity is no longer a basic law. It may become a useful approximation, but not a basic law. Once a fundamental law like the principle of relativity is discovered or accepted, it has a profound effect on the way we think about things. In this case, if there is no way that we can detect our own uniform motion, then we might as well ignore our motion and always assume that we are at rest. Nature is usually easier to explain if we take the point of view that we are at rest and that other people and things are moving by. It is the principle of relativity that allows us to take this self-centered point of view.

It is a shock, a lot of excitement is generated, when what was accepted as a basic law of physics is discovered not to be exactly true. The discovery usually occurs in some obscure corner of science where no one thought to look before. And it will probably have little effect on most practical applications. But the failure of a basic law changes the way we think.

Suppose, for example, that it was discovered that the principle of relativity did not apply to the decay of an esoteric elementary particle created only in the gigantic particle accelerating machines physicists have recently built. This violation of the principle of relativity would have no practical effect on our daily lives, but it would have a profound psychological effect. We would then know that our uniform motion could be detected, and therefore on a fundamental basis we could no longer take the point of view that we are at rest and others are moving. There would be legitimate debates as to who was moving and who was at rest. We would search for a formulation of the laws of physics that made it intuitively clear who was moving and who was at rest.

This is almost what happened in 1860. In that year, James Clerk Maxwell summarized the laws of electricity and magnetism in four short equations. He then solved these equations to predict the existence of a wave of electric and magnetic force that should travel at a speed of approximately  $3 \times 10^8$  meters per second. The predicted speed, which we will call c, could be determined from simple measurements of the behavior of an electric circuit.

Before Maxwell, no one had considered the possibility that electric and magnetic forces could combine in a wavelike structure that could travel through space. The first question Maxwell had to answer was what this wave was. Did it really exist? Or was it some spurious solution of his equations? The clue was that the speed c of this wave was so fast that only light had a comparable speed. And more remarkably the known speed of light, and the speed c of his wave were very close—to within experimental error they were equal. As a consequence Maxwell proposed that he had discovered the theory of light, and that this wave of electric and magnetic force was light itself.

Maxwell's theory explained properties of light such as polarization, and made predictions like the existence of radio waves. Many predictions were soon verified, and within a few years there was little doubt that Maxwell had discovered the theory of light.

One problem with Maxwell's theory is that by measurements of the speed of light, it appears that one should be able to detect one's own uniform motion. In the next section we shall see why. This had two immediate consequences. One was a change in the view of nature to make it easy to see who was moving and who was not. The second was a series of experiments to see if the earth were moving or not.

In the resulting view of nature, all of space was filled with an invisible substance called *ether*. Light was pictured as a wave in the ether medium just as ocean waves are waves in the medium of water. The experiments, initiated by Michaelson and Morley, were designed to detect the motion of the earth by measuring how fast the earth was moving through the ether medium.

The problem with the ether theory was that all experiments designed to detect ether, or to detect motion through it, seemed to fail. The more clever the experiment, the more subtle the apparent reason for the failure. We will not engage in any further discussion of the ether theory, because ether still has never been detected. But we will take a serious look in the next section at how the measurement of the speed of a pulse of light should allow us to detect our own uniform motion. And then in the rest of the chapter we will discuss how a young physicist, working in a patent office in 1905, handled the problem.



#### Figure 2

Rain drops creating circular waves on the surface of a puddle. (Courtesy Bill Jack Rodgers, Los Alamos Scientific Laboratory.)

# WAVE MOTION

We do not need to know the details of Maxwell's theory to appreciate how one should be able to use the theory to violate the principle of relativity. All we need is an understanding of some of the basic properties of wave motion.

The most familiar examples of wave motion are the waves on the surface of water. We have seen the waves that spread out in circles when a stone is dropped in a pond, or rain hits a puddle in a sidewalk as shown in Figure (2). And most of us have seen the ocean waves destroying themselves as they crash into the beach. The larger ocean waves often originate at a storm far out to sea, and have traveled hundreds or even a thousand miles to reach you (see Figure (3)). The very largest ocean waves, created by earthquakes or exploding volcanos have been known to travel almost around the earth.



**Figure 3** This ocean wave traveled hundreds of miles from Hurricane Bertha to the Maine coast (July 31, 1990).

Although you cannot see them, sound waves are a more familiar form of wave motion. Sound moving through air, waves moving over water, and light, all have certain common features and ways of behaving which we classify as *wave motion*. In later chapters we will study the subject of wave motion in considerable detail. For now we will limit our discussion to a few of the features we need to understand the impact of Maxwell's theory.

Two examples of wave motion that are easy to study are a wave pulse traveling down a rope as indicated in Figure (4) or down a stretched Slinky® (the toy coil that 'climbs' down stairs) as shown in Figure (5). The advantage of using a stretched Slinky is that the waves travel so slowly that you can study them as they move. It turns out that the speed of a wave pulse depends upon the medium along which, or through which, it is traveling. For example, the speed of a wave pulse along a rope or Slinky is given by the formula



**Figure 4** *Wave pulse traveling along a rope.* 



**Figure 5** *Wave pulse traveling along a Slinky.* 

Speed of  
wave pulse = 
$$\sqrt{\frac{\tau}{\mu}}$$
 (1)

where  $\tau$  is the tension in the rope or Slinky, and  $\mu$  the mass per unit length. Do not worry about precise definitions of tension or mass, the important point is that there is a formula for the speed of the wave pulse, a formula that depends only on the properties of the medium along which the pulse is moving.

The speed does not depend upon the shape of the pulse or how the pulse was created. For example, the Slinky pulse travels much more slowly than the pulse on the rope because the suspended Slinky has very little tension  $\tau$ . We can slow the Slinky wave down even more by hanging crumpled pieces of lead on each end of the coils of the Slinky to increase its mass per unit length  $\mu$ .

Another kind of wave we can create in the Slinky is the so called *compressional wave* shown in Figure (6). Here the end of the Slinky was pulled back and released, giving a moving pulse of compressed coils. The formula for the speed of the compression wave is still given by Equation (1), if we interpret  $\tau$  as the stiffness (*Youngs modulus*) of the suspended Slinky.

If we use a loudspeaker to produce a compressional pulse in air, we get a sound wave that travels out from the loudspeaker at the speed of sound. The formula for the speed of a sound wave is

$$\frac{\text{Speed of}}{\text{sound}} = \sqrt{\frac{B}{\rho}}$$
(2)

where B is the *bulk modulus* which can be thought of as the rigidness of the material, and the mass per unit length  $\mu$  is replaced by the mass per unit volume  $\rho$ .



#### Figure 6

To create a compressional wave on a suspended Slinky, pull the end back a bit and let go. A substance like air, which is relatively compressible, has a small rigidity B, while substances like steel and granite are very rigid and have large values of B. As a result sound travels much faster in steel and granite than in air. For air at room temperature and one atmosphere of pressure, the speed of sound is 343 meters or 1125 feet per second. Sound travels about 20 times faster in steel and granite. Again the important point is that the speed of a wave depends on the properties of the medium through which it is moving, and not on the shape of the wave or the way it was produced.

# **Measurement of the Speed of Waves**

If you want to know how fast your car is traveling you look at the speedometer. Some unknown machinery in the car makes the needle of the speedometer point at the correct speed. Since the wave pulses we are discussing do not have speedometers, we have to carry out a series of measurements in order to determine their speed. In this section we wish to discuss precisely how the measurements can be made using meter sticks and clocks so that there will be no ambiguity, no doubt about precisely what we mean when we talk about the speed of a wave pulse. We will use the Slinky wave pulse as our example, because the wave travels slowly enough to actually carry out these measurements in a classroom demonstration.

The first experiment, shown in Figure (7), involves two students and the instructor. One student stands at the end of the stretched Slinky and releases a wave pulse like that shown in Figure (6). The instructor holds a meter stick up beside the Slinky as shown. The other student has a stopwatch and measures the length of time it takes the pulse to travel from the front to the back of the stick. (She presses the button once when the pulse reaches the front of the meter stick, presses it again when the pulse gets to the back, and reads the elapsed time T.) The speed of the pulse is then defined to be

Speed of  
Slinky pulse = 
$$\frac{1 \text{ meter}}{T \text{ seconds}}$$
 (3)





Later in the course, when we have discussed ways of measuring tension  $\tau$  and mass per unit length  $\mu$ , we can compare the experimental result we get from Equation (3) with the theoretically predicted result of Equation (1). With a little practice using the stopwatch, it is not difficult to get reasonable agreement between theory and experiment.

In our second experiment, shown in Figure (8a) everything is the same except that the instructor has been replaced by a student, let us say it is Bill, holding the meter stick and running toward the student who releases the wave pulse. Again the second student measures the length of time it takes the pulse to travel from the front to the back of the meter stick. Let us call



**Figure 8a** Bill runs toward the source of the pulse while measuring its speed.

this the time  $T_1$ . This time  $T_1$  is less than T because Bill and the meter stick are moving toward the pulse.

To Bill, the pulse passes his one meter long stick in a time  $T_1$ , therefore the speed of the pulse past him is

$$v_1 = \frac{1 \text{ meter}}{T_1 \text{ seconds}} = \frac{\text{speed of pulse}}{\text{relative to Bill}}$$
 (4a)

Bill should also have carried the stopwatch so that  $v_1$  would truly represent his measurement of the speed of the pulse. But it is too awkward to hold the meter stick, and run and observe when the pulse is passing the ends of the stick.

The speed  $v_1$  measured by Bill is not the same as the speed v measured by the instructor in Figure (7).  $v_1$  is greater than v because Bill is moving toward the wave pulse. This is not surprising: if you are on a freeway and everyone is traveling at a speed v = 55 miles per hour, the oncoming traffic in the opposite lane is traveling past you at a speed of 110 miles per hour because you are moving toward them.

In Figure (8b) we again have the same situation as in Figure (7) except that Bill is now replaced by Joan who is running away from the student who releases the pulse. Joan is moving in the same direction as the pulse and it takes a longer time  $T_2$  for the pulse to pass her. (Assume that Joan is not running faster than the pulse.) The speed of the pulse relative to Joan is

$$v_2 = \frac{1 \text{ meter}}{T_2 \text{ seconds}} = \frac{\text{speed of pulse}}{\text{relative to Joan}}$$
 (4b)

Joan's speed  $v_2$  will be considerably less than the speed v observed by the instructor.



**Figure 8b** Joan runs away from the source of the pulse while measuring its speed.

In these three experiments, the instructor is special (wouldn't you know it). Only the instructor measures the speed v predicted by theory, only for the instructor is the speed given by  $v = \sqrt{\tau/\mu}$ . Both the students Bill and Joan observe different speeds, one larger and one smaller than the theoretical value.

What is special about the instructor? In this case the instructor gets the predicted answer because she is at rest relative to the Slinky. If we hadn't seen the experiment, but just looked at the answers, we could tell that the instructor was at rest because her result agreed with the predicted speed of a Slinky wave. Bill got too high a value because he was moving toward the pulse; Joan, too low a value because she was moving in the same direction.

The above set of experiments is not strikingly profound. In a sense, we have developed a new and rather cumbersome way to tell who is not moving relative to the Slinky. But the same procedures can be applied to a series of experiments that gives more interesting results. In the new series of experiments, we will use a pulse of light rather than a wave pulse on a Slinky.

Since the equipment is not likely to be available among the standard set of demonstration apparatus, and since it will be difficult to run at speeds comparable to the speed of light we will do this as a thought experiment. We will imagine that we can measure the time it takes a light pulse to go from the front to the back of a meter stick. We will imagine the kind of results we expect to



#### Figure 9

Apparatus for the thought experiment. Now we wish to measure the speed of a laser wave pulse, rather than the speed of a Slinky wave pulse. The photo detectors are used to measure the length of time the laser pulse takes to pass by the meter stick. get, and then see what the consequences would be if we actually got those results.

The apparatus for our new thought experiment is shown in Figure (9). We have a laser which can produce a very short pulse of light – only a few millimeters long. The meter stick now has photo detectors and clocks mounted on each end, so that we can accurately record the times at which the pulse of light passed each end. These clocks were synchronized, so the time difference is the length of time T it takes the pulse of light to pass the meter stick.

Before the experiment, the instructor gives a short lecture to the class. She points out that according to Maxwell's theory of light, a light wave should travel at a speed c given by the formula

$$c = \frac{1}{\sqrt{\mu_0 \varepsilon_0}} \tag{5}$$

where  $\mu_0$  and  $\varepsilon_0$  are constants in the theory of electricity. She says that later on in the year, the students will perform an experiment in which they measure the value of the product  $\mu_0\varepsilon_0$ . This experiment involves measuring the size of coils of wire and plates of aluminum, and timing the oscillation of an electric current sloshing back and forth between the plates and the coil. The important point is that these measurements do not involve light. It is analogous to the Slinky where the predicted speed  $\sqrt{\tau/\mu}$  of a Slinky wave involved measurements of the stiffness  $\tau$  and mass per unit length  $\mu$ , and had nothing to do with observations of a Slinky wave pulse.



Figure 10a Plates and coil for measuring the experimental value of  $\mu_0 \varepsilon_0$ .



**Figure 10b** *The plates and coil we use in the laboratory.* 





Although she is giving out the answer to the lab experiment, she points out that the value of c from these measurements is

$$c = \frac{1}{\sqrt{\mu_0 \varepsilon_0}} = 3 \times 10^8 \text{ meters/second}$$
 (6)

which is a well-known but uncomfortably large and hard to remember number. However, she points out,  $3 \times 10^8$  meters is almost exactly one billion (10<sup>9</sup>) feet. If you measure time, not in seconds, but in billionths of a second, or *nanoseconds*, where

$$1 \text{ nanosecond} \equiv 10^{-9} \text{ seconds} \tag{7}$$

laser

then since light travels only one foot in a nanosecond, the speed of light is simply



**Figure 12a** Bill runs toward the source of the pulse while measuring its speed.

$$c = 3 \times 10^{8} \frac{\text{meter}}{\text{sec}} = 10^{9} \frac{\text{feet}}{\text{sec}}$$

$$c = \frac{1 \text{ foot}}{10^{-9} \text{sec}}$$
(8)
$$c = 1 \frac{\text{foot}}{\text{nanosecond}}$$

She says that because this is such an easy number to remember, she will use it throughout the rest of the course.

The lecture on Maxwell's theory being over, the instructor starts in on the thought experiment. In the first run she stands still, holding the meter stick, and the student with the laser emits a pulse of light as shown in Figure (11). The pulse passes the 3.28 foot length of the meter stick in an elapsed time of 3.28 nanoseconds, for a measured speed

$$v(\text{light pulse}) = \frac{3.28 \text{ feet}}{3.28 \text{ nanoseconds}}$$
$$= 1 \frac{\text{foot}}{\text{nanosecond}}$$
(9)

The teacher notes, with a bit of complacency, that she got the predicted speed of 1 foot/nanosecond. Again, the instructor is special.

Then the instructor invites Bill to hold the meter stick and run toward the laser as shown in Figure (12a). Since this is a thought experiment, she asks Bill to run at nearly the speed of light, so that the time should be cut in half and Bill should see light pass him at nearly a speed of 2c.





**Figure 12b** Joan runs away from the source of the pulse while measuring its speed.

Then she invites Joan to hold the meter stick and run at about half the speed of light in the other direction as shown in Figure (12b). One would expect that the light would take twice as long to pass Joan as it did the instructor and that Joan should obtain a value of about c/2 for the speed of light.

Suppose it turned out this way. Suppose that the instructor got the predicted answer 1 foot/nanosecond, while Bill who is running toward the pulse got a higher value and Joan, running with the pulse got a lower value. Just as in our Slinky pulse experiment we could say that the instructor was at rest while both Bill and Joan were moving.

*But, moving relative to what?* In the Slinky experiment, the instructor was at rest relative to the Slinky – the medium through which a Slinky wave moves. Light pulses travel through empty space. Light comes to us from stars 10 billion light years away, almost across the entire universe. The medium through which light moves is empty space.

If the experiment came out the way we described, the instructor would have determined that she was at rest relative to empty space, while Bill and Joan would have determined that they were moving. *They would have violated the principle of relativity*, which says that you cannot detect your own motion relative to empty space.

The alert student might argue that the pulses of light come out of the laser like bullets from a gun at a definite muzzle velocity, and that all the instructor, Bill and Joan are doing is measuring their speed relative to the laser. Experiments have carefully demonstrated that the speed of a pulse of light depends in no way on the motion of the emitter just as the Slinky pulse depended in no way on how the student started the pulse. Maxwell's theory predicts that light is a wave, and many experiments have verified the wave nature of light, including the fact that its speed does not depend on how it was emitted.

From the logical simplicity of the above thought experiment, from the ease with which we should be able to violate the principle of relativity (if we could accurately measure the speed of a pulse of light passing us), it is not surprising that after Maxwell developed this theory of light, physicists did not take the principle of relativity seriously, at least for the next 45 years.

# **Michaelson-Morley Experiment**

The period from 1860 to 1905 saw a number of attempts to detect one's own or the earth's motion through space by measuring the speed of pulses of light. Actually it was easier and far more accurate to compare the speeds of light traveling in different directions. If you were moving forward through space (like Bill in our thought experiment), you should see light coming from in front of you traveling faster than light from behind or even from the side.

Michaelson and Morley used a device called a Michaelson interferometer which compared the speeds of pulses of light traveling at right angles to each other. A detailed analysis of their device is not hard, just a bit lengthy. But the result was that the device should be able to detect small differences in speeds, small enough differences so that the motion of the earth through space should be observable -- even the motion caused by the earth orbiting the sun.

At this point we can summarize volumes of the history of science by pointing out that no experiment using the Michaelson interferometer, or any device based on measuring or comparing the speed of light pulses, ever succeeded in detecting the motion of the earth.

#### **Exercise 3**

Units of time we will often use in this course are the millisecond, the microsecond, and the nanosecond, where

1 millisecond =  $10^{-3}$  seconds (one thousandth) 1 microsecond =  $10^{-6}$  seconds (one millionth) 1 nanosecond =  $10^{-9}$  seconds (one billionth)

How many feet does light travel in

- a) one millisecond (1ms)?
- b) one microsecond  $(1\mu s)$ ?
- c) one nanosecond (1ns)?

# EINSTEIN'S PRINCIPLE OF RELATIVITY

In 1905 Albert Einstein provided a new perspective on the problems we have been discussing. He was apparently unaware of the Michaelson-Morley experiments. Instead, Einstein was familiar with Maxwell's equations for electricity and magnetism, and noted that these equations had a far simpler form if you took the point of view that you are at rest. He suggested that these equations took this simple form, not just for some privileged observer, but for everybody. If the principle of relativity were correct after all, then everyone, no matter how they were moving, could take the point of view that they were at rest and use the simple form of Maxwell's equations.

How did Einstein deal with measurements of the speed of light? We have seen that if someone, like Bill in our thought experiment, detects a pulse of light coming at them at a speed faster than c = 1 foot/nanosecond, then that person could conclude that they themselves were moving in the direction from which the light was coming. They would have thereby violated the principle of relativity.

Einstein's solution to that problem was simple. He noted that any measurement of the speed of a pulse of light that gave an answer different from c = 1 foot/ nanosecond could be used to violate the principle of relativity. Thus if the principle of relativity were correct, *all measurements of the speed of light must give the answer c*.

Let us put this in terms of our thought experiment. Suppose the instructor observed that the light pulse passed the 3.24 foot long meter stick in precisely 3.24 nanoseconds. And suppose that Bill, moving at nearly the speed of light toward the laser, also observed that the light took 3.24 nanoseconds to pass by his meter stick. And suppose that Joan, moving away from the laser at half the speed of light, also observed that the pulse of light took 3.24 nanoseconds to pass by her meter stick. If the instructor, Bill and Joan all got precisely the same answer for the speed of light, then none of their results could be used to prove that one was at rest and the others moving. Since their answer of 3.24 feet in 3.24 nanoseconds or 1 foot/nanosecond is in agreement with the predicted value  $c = 1/\sqrt{\mu_0\epsilon_0}$ 

from Maxwell's theory, they could all safely assume that they were at rest. At the very least, their measurements of the speed of the light pulse could not be used to detect their own motion.

As we said, the idea is simple. You always get the answer c whenever you measure the speed of a light pulse moving past you. But the idea is horrendous. Einstein went against more than 200 years of physics and centuries of observation with this suggestion.

Suppose, for example, we heard about a freeway where all cars traveled at precisely 55 miles per hour – **no exceptions**. Hearing about this freeway, our three people in the thought experiment decide to test the rule. The instructor sets up measuring equipment in the median strip and observes that the rule is correct. Cars in the north bound lane travel north at 55 miles per hour, and cars in the south bound lane go south at 55 miles per hour.

For his part of the experiment, Bill gets into one of the north bound cars. Since Bill knows about the principle of relativity he takes the point of view that he is at rest. If the 55 miles per hour speed is truly a fundamental law, then he, who is at rest, should see the south bound cars pass at 55 miles per hour.

Likewise, Joan, who is in a south bound car, can take the point of view that she is at rest. She knows that if the 55 miles per hour speed limit is a fundamental law, then north bound cars must pass her at precisely 55 miles per hour. If the instructor, Bill and Joan all observe that every car on the freeway always passes them at the same speed of 55 miles per hour, then none of them can use this observation to detect their own motion.

Freeways do not work that way. Bill will see south bound cars passing him at 110 miles per hour. And Joan will see north bound cars passing at 110 miles per hour. From these observations Bill and Joan will conclude that in fact they are moving – at least relative to the freeway.

Measurements of the speed of a pulse of light differ, however, in two significant ways from measurements of the speed of a car on a freeway. First of all, light moves through empty space, not relative to anything. Secondly, light moves at enormous speeds, speeds that lie completely outside the realm of common experience. Perhaps, just perhaps, the rules we have learned so well from common experience, do not apply to this realm. The great discoveries in physics often came when we look in some new realm on the very large scale, or the very small scale, or in this case on the scale of very large, unfamiliar speeds.

# The Special Theory of Relativity

Einstein developed his special theory of relativity from two assumptions:

# 1) The principle of relativity is correct.

# 2) Maxwell's theory of light is correct.

As we have seen, the only way Maxwell's theory of light can be correct and not violate the principle of relativity, is that every observer who measures the speed of light, must get the predicted answer  $c = 1/\sqrt{\mu_0 \epsilon_0} = 1$  foot/nanosecond. Temporarily we will use this as the statement of Einstein's second postulate:

# 2a) Everyone, no matter how he or she is moving, must observe that light passes them at precisely the speed c.

Postulates (1) and (2a) salvage both the principle of relativity and Maxwell's theory, but what else do they predict? We have seen that measurements of the speed of a pulse of light do not behave in the same way as measurements of the speed of cars on a freeway. Something peculiar seems to be happening at speeds near the speed of light. What are these peculiar things? How do we find out?

To determine the consequences of his two postulates, Einstein borrowed a technique from Galileo and used a series of thought experiments. Einstein did this so clearly, explained the consequences so well in his 1905 paper, that we will follow essentially the same line of reasoning. The main difference is that Einstein made a number of strange predictions that in 1905 were hard to believe. But these predictions were not only verified, they became the cornerstone of much of 20th century physics. We will be able to cite numerous tests of all the predictions.

# **Moving Clocks**

Our first thought experiment for Einstein's special relativity will deal with the behavior of clocks. We saw that the measurement of the speed of a pulse of light required a timing device, and perhaps the peculiar results can be explained by the peculiar behavior of the timing device.

Also the peculiar behavior seems to happen at high speeds near the speed of light, not down at freeway speeds. Thus the question we would like to ask is what happens to a clock that is moving at a high speed, near the speed of light?

That is a tough question. There are many kinds of clocks, ranging from hour glasses dripping sand, to the popular digital quartz watches, to the atomic clocks used by the National Bureau of Standards. The oldest clock, from which we derive our unit of time, is the motion of the earth on its axis each 24 hours. We have both the problem of deciding which kind of clock we wish to consider moving at high speeds, and then figure out how that clock behaves.

The secret of working with thought experiments is to keep everything as simple as possible and do not try to do too much at once. If we want to understand what happens to a moving clock, we should start with the simplest clock we can find. If we cannot understand that one, we will imagine an even simpler one.

A clock that is fairly easy to understand is the old grandfather's clock shown in Figure (13), where the timing device is the swinging pendulum. There are also wheels, gears, and hands, but these merely count swings of the pendulum. The pendulum itself is what is important. If you shorten the pendulum it swings faster and the hands go around faster.



**Figure 13** *Grandfather's clock.* 

We could ask what we would see if we observed a grandfather's clock moving past us at a high speed, near the speed of light. The answer is likely to be "I don't know". The grandfather's clock, with its swinging pendulum mechanism, is still too complicated.

A simpler timing device was considered by Einstein, namely a bouncing pulse of light. Suppose, we took the grandfather's clock of Figure (13), and replaced the pendulum by two mirrors and a pulse of light as shown in Figure (14). Space the mirrors 1 foot apart so that the pulse of light will take precisely one nanosecond to bounce either up or down. Leave the rest of the machinery of the grandfather's clock more or less intact. In other words have the wheels and gears now count bounces of the pulse of light rather than swings of the pendulum. And recalibrate the face of the clock so that for each bounce, the hand advances one nanosecond. (The marvelous thing about thought experiments is that you can get away with this. You do not have to worry about technical feasibility, only logical consistency.)

The advantage of replacing the pendulum with a bouncing light pulse is that, so far, the only thing whose behavior we understand when moving at nearly the speed of light is light itself. We know that light always moves at the speed c in all circumstances, to any observer. If we use a bouncing light pulse as a timing device, and can figure out how the pulse behaves, then we can figure out how the clock behaves.



## Figure 14

Light pulse clock. We can construct a clock by having a pulse of light bounce between two mirrors. If the mirrors are one foot apart, then the time between bounces will be one nanosecond. The face of the clock displays the number of bounces.

For our thought experiment it is convenient to construct two identical light pulse clocks as shown in Figure (15). We wish to take great care that they are identical, or at least that they run at precisely the same rate. Once they are finished, we adjust them so that the pulses bounce up and down together for weeks on end.

Now we get to the really hypothetical part of our thought experiment. We give one of the clocks to an astronaut, and we keep the other for reference. The astronaut is instructed to carefully pack his clock, accelerate up to nearly the speed of light, unpack his clock, and go by us at a constant speed so that we can compare our reference clock to his moving clock.

Before we describe what we see, let us take a look at a brief summary of the astronaut's log book of the trip. The astronaut writes, "Icarefully packed the light pulse clock because I did not want it damaged during the accelerations. My ship can maintain an acceleration of 5gs, and even then it took about a month to get up to our final speed of just over half the speed of light."

"Once the accelerations were over and I was coasting, I took the light pulse clock out of its packing and set it up beside the window, so that the class could see the clock as it went by. Before the trip I was worried that I might have some trouble getting the light pulse into the clock, but it was no problem at all. I couldn't even tell that I was moving! The light pulse went in and the clock started ticking just the way it did back in the lab, before we started the trip."

"It was not long after I started coasting, that the class went by. After that, I packed everything up again, decelerated, and returned to earth."



**Figure 15** *Two identical light pulse clocks.* 



#### Figure 16

In order to stay in the astronaut's moving clock, the light pulse must follow a longer, saw-tooth, path.

What we saw as the astronaut went by is illustrated in the sketch of Figure (16). On the left is our reference clock, on the right the astronaut's clock moving by. You will recall that the astronaut had no difficulty getting the light pulse to bounce, and as a result we saw his clock go by with the pulse bouncing inside.

For his pulse to stay in his clock, his pulse had to travel along the saw-tooth path shown in Figure (16). The saw-tooth path is longer than the up and down path taken by the pulse in our reference clock. His pulse had to travel farther than our pulse to tick off one nanosecond.

Here is what is peculiar. If Einstein's postulate is right, if the speed of a pulse of light is always c under any circumstances, then our pulse bouncing up and down,

and the astronaut's pulse traveling along the saw tooth path are both traveling at the same speed c. Since the astronaut's pulse travels farther, the astronaut's clock must take longer to tick off a nanosecond. *The astronaut's clock must be running slower!* 

Because there are no budget constraints in a thought experiment, we are able to get a better understanding of how the astronaut's clock was behaving by having the astronaut repeat the trip, this time going faster, about .95 c. What we saw is shown in Figure (17). The astronaut's clock is moving so fast that the saw tooth path is stretched way out. The astronaut's pulse takes a long time to climb from the bottom to the top mirror in his clock, his nanoseconds take a long time, and his clock runs very slowly.



#### Figure 17

When the astronaut goes faster, his light pulse has to go farther in order to register a bounce. Since the speed of light does not change, it takes longer for one bounce to register, and the astronaut's moving clock runs slower.

It does not take too much imagination to see that if the astronaut came by at the speed of light c, the light pulse, also traveling at a speed c, would have to go straight ahead just to stay in the clock. It would never be able to get from the bottom to the top mirror, and his clock would never tick off a nanosecond. His clock would stop!

#### **Exercise 4**

Discuss what the astronaut should have seen when the class of students went by. In particular, draw the astronaut's version of Figure (16) and describe the situation from the astronaut's point of view.

It is not particularly difficult to calculate the amount by which the astronaut's clock runs slow. All that is required is the Pythagorean theorem. In Figure (18), on the left, we show the path of the light pulse in our reference clock, and on the right the path in the astronaut's moving clock. Let T be the length of time it takes our pulse to go from the bottom to the top mirror, and T' the longer time light takes to travel along the diagonal line from his bottom mirror to his top mirror. We can think of T as the length of one of our nanoseconds, and T' as the length of one of the astronaut's longer nanoseconds.

The distance an object, moving at a speed v, travels in a time T, is vT. (If you go 30 miles per hour for 3 hours, you travel 90 miles.) Thus, in Figure (18), the distance our light pulse travels in going from the bottom to the top mirror is cT as shown. The astronaut's light pulse, which takes a time T' to travel the diagonal path, must have gone a distance cT' as shown.

During the time T', while the astronaut's light pulse is going along the diagonal path, the astronaut's clock, which is traveling at a speed v, moves forward a distance vT' as shown. This gives us a right triangle whose base is vT', whose hypotenuse is cT', and whose height, determined from our clock, is cT. According to the Pythagorean theorem, these sides are related by

$$(cT')^2 = (vT')^2 + (cT)^2$$
 (10)

Carrying out the squares, and collecting the terms with T' on one side, we get

$$c^{2}T'^{2} - v^{2}T'^{2} = (c^{2} - v^{2})T'^{2} = c^{2}T^{2}$$
$$T'^{2} = \frac{c^{2}T^{2}}{c^{2} - v^{2}} = \frac{c^{2}T^{2} \times (1/c^{2})}{(c^{2} - v^{2}) \times (1/c^{2})} = \frac{T^{2}}{1 - v^{2}/c^{2}}$$

Taking the square root of both sides gives

$$T' = \frac{T}{\sqrt{1 - v^2/c^2}}$$
(11)

Equation (11) gives a precise relationship between the length of our nanosecond T and the astronaut's longer nanosecond T'. We see that the astronaut's basic time unit T' is longer than our basic time unit T by a factor  $1/\sqrt{1 - v^2/c^2}$ .

The factor  $1/\sqrt{1 - v^2/c^2}$  appears in a number of calculations involving Einstein's special theory of relativity. As a result, it is essential to develop an intuitive feeling for this number. Let us consider several examples to begin to build this intuition. If v = 0, then

$$T' = \frac{T}{\sqrt{1 - v^2/c^2}} = \frac{T}{\sqrt{1 - 0}}$$
$$T' = \frac{T}{1} = T \qquad (v = 0)$$
(12)



#### Figure 18

In our clock, the light pulse travels a distance cT in one bounce. In the astronaut's clock, the pulse travels a distance cT' while the clock moves forward a distance vT' during one bounce. and we see that a clock at rest keeps the same time as ours. If the astronaut goes by at one tenth the speed of light, v = .1 c, and we get

$$T' = \frac{T}{\sqrt{1 - (.1c)^2/c^2}} = \frac{T}{\sqrt{1 - .01}}$$
$$T' = \frac{T}{\sqrt{.99}} = 1.005T \quad (v = c/10) \quad (13)$$

In this case the astronaut's seconds lengthen only by a factor 1.005 which represents only a .5% increase. If the astronaut's speed is increased to half the speed of light, we get

$$T' = \frac{T}{\sqrt{1 - (.5c)^2/c^2}} = \frac{T}{\sqrt{1 - .25}}$$
$$T' = \frac{T}{\sqrt{.75}} = 1.15T \quad (v = c/2) \quad (14)$$

Now we are getting a 15% increase in the length of the astronaut's seconds.

When we work with atomic or subatomic particles, it is not difficult to accelerate these particles to speeds close to the speed of light. Shortly we will consider a particle called a muon, that is traveling at a speed v = .994 c. For this particle we have

$$T' = \frac{T}{\sqrt{1 - (.994c)^2/c^2}} = \frac{T}{\sqrt{1 - .988}}$$

$$T' = \frac{T}{\sqrt{.012}} = 9T \quad (v = .994c)$$
(15)

Here we are beginning to see some large effects. If the astronaut were traveling this fast, his seconds would be 9 times longer than ours, his clock would be running only 1/9th as fast.

If we go all the way to v = c, Equation (11) gives

$$T' = \frac{T}{\sqrt{1 - c^2/c^2}} = \frac{T}{\sqrt{1 - 1}}$$
$$T' = \frac{T}{\sqrt{0}} = \infty \qquad (v = c) \qquad (16)$$

In this case, the astronaut's seconds would be infinitely long and the astronaut's clock would stop. This agrees with our earlier observation that if the astronaut went by at the speed of light, the light pulse in his clock would have to go straight ahead just to stay in the clock. It would not have time to move up or down, and therefore not be able to tick off any seconds.

So far we have been able to use a pocket calculator to evaluate  $1/\sqrt{1 - v^2/c^2}$ . But if the astronaut were flying in a commercial jet plane at a speed of 500 miles per hour, you have problems because  $1/\sqrt{1 - v^2/c^2}$  is so close to 1 that the calculator cannot tell the difference. In a little while we will show you how to do such calculations, but for now we will just state the answer.

$$\frac{1}{\sqrt{1 - v^2/c^2}} = 1 + 2.7 \times 10^{-13} \frac{\text{for a speed}}{\text{of 500 mi/hr}}$$
(17)

To put this result in perspective suppose the astronaut flew on the jet for what we thought was a time T = 1 hour or 3600 seconds. The astronaut's light pulse clock would show a longer time T' given by

$$T' = \frac{T}{\sqrt{1 - v^2/c^2}}$$
$$= (1 + 2.7 \times 10^{-13}) \times (3600 \text{ seconds})$$
$$= 1 \text{ hour } + .97 \times 10^{-9} \text{ seconds}$$

Since  $.97 \times 10^{-9}$  seconds is close to a nanosecond, we can write

$$T' \approx 1 \text{ hour} + 1 \text{ nanosecond}$$
 (18)

The astronaut's clock takes 1 hour plus 1 nanosecond to move its hand forward 1 hour. We would say that his light pulse clock is losing a nanosecond per hour.

Students have a tendency to memorize formulas, and Equation (11),  $T' = T/\sqrt{1 - v^2/c^2}$  looks like a good candidate. **But don't!** If you memorize this formula, you will mix up T' and T, forgetting which seconds belong to whom. **There is a much easier way to always** get the right answer.

For any speed v less than or equal to c (which is all we will need to consider) the quantity  $\sqrt{1 - v^2/c^2}$  is always a number less than or equal to 1, and  $1/\sqrt{1 - v^2/c^2}$  is always greater than or equal to 1. For the examples we have considered so far, we have

V	$\sqrt{1 - \mathbf{v}^2/\mathbf{c}^2}$	$1/\sqrt{1 - v^2/c^2}$
0	1	1
500 mi/hr	$1 - 2.7 \times 10^{-13}$	$1 + 2.7 \times 10^{-13}$
c/10	.995	1.005
c/2	.87	1.15
.994c	1/9	9
С	0	∞
	•	1

# Table 1

You also know intuitively that for the moving light pulse clock, the light pulse travels a longer path, and therefore the moving clock's seconds are longer.

If you remember that  $\sqrt{1 - v^2/c^2}$  appears somewhere in the formula, all you have to do is ask yourself what to do with a number less than one to make the answer bigger; clearly, you have to divide by it.

As an example of this way of reasoning, note that if a moving clock's seconds are longer, then the rate of the clock is slower. The number of ticks per unit time is less. If we want to talk about the rate of a moving clock, do we multiply or divide by  $\sqrt{1 - v^2/c^2}$ ? To get a reduced rate, we multiply by  $\sqrt{1 - v^2/c^2}$  since that number is always less than one. Thus we can say that

the rate of the moving clock is reduced by a factor  $\sqrt{1 - v^2/c^2}$ .

The factor  $\sqrt{1 - v^2/c^2}$  will appear numerous times throughout the text. But in every case you should have an intuitive idea of whether the quantity under consideration should increase or decrease. If it increases, divide by the  $\sqrt{1 - v^2/c^2}$ , and if it decreases, multiply by  $\sqrt{1 - v^2/c^2}$ . This approach gives the right answer, reduces memorization, and eliminates obscure notation like T' and T.

# **Other Clocks**

So far we have an interesting but limited result. We have predicted that if someone carrying a light pulse clock moves by us at a speed v, we will see that their light pulse clock runs slow by a factor  $\sqrt{1 - v^2/c^2}$ . Up until now we have said nothing about any other kind of clock, and we have the problem that no one has actually constructed a light pulse clock. But we can easily generalize our result with another thought experiment fairly similar to the one we just did.

For the new thought experiment let us rejoin the discussion between the astronaut and the class of students. We begin just after the students have told the astronaut what they saw. "I was afraid of that," the astronaut replies. "I never did trust that light pulse clock. I am not at all surprised that it ran slow. But now my digital watch, it's really good. It is based on a quartz crystal and keeps really good time. It wouldn't run slow like the light pulse clock."

"I'll bet it would," Bill interrupts. "How much?," the astronaut responds indignantly. "The cost of one more trip," Bill answers.

In the new trip, the astronaut is to place his digital watch right next to the light pulse clock so that the astronaut and the class can see both the digital watch and the light pulse clock at the same time. The idea is to compare the rates of the two timing devices.

"Look what would happen," Bill continues, "if your digital watch did not slow down. When you come by, your digital watch would be keeping "God's time" as you call it, while your light pulse clock would be running slower."

"The important part of this experiment is that because the faces of the two clocks are together, if we see them running at different rates, you will too. You would notice that here on earth, when you are at rest, the two clocks ran at the same rate. But when you were moving at high speed, they would run at different rates. You could use this difference in rates to detect your own motion, and therefore violate the principle of relativity."

The astronaut thought about this for a bit, and then responded, "I'll grant that you are partly right. On my previous trips, after the accelerations ceased and I started coasting toward the class, I did not feel any motion. I had no trouble unpacking the equipment and setting it up. The light pulse went in just as it had back in the lab, and I was sure that the light pulse clock was working just fine. I certainly would have noticed any difference in the rates of the two clocks." "Are you insinuating," the astronaut continued, "that the reason I did not detect my light pulse running slow was because my digital watch was also running slow?"

"Almost," replied Bill, "but you have other timing devices in your capsule. You shave once a day because you do not like the feel of a beard. This is a cyclic process that could be used as the basis of a new kind of clock. If your shaving cycle clock did not slow down just like the light pulse clock, you could time your shaving cycle with the light pulse clock and detect your motion. You would notice that you had to shave more times per light pulse month when you were moving than when you were at rest. This would violate the principle of relativity."

"Wow," the astronaut exclaimed, "if the principle of relativity is correct, and the light pulse clock runs slow, then every process, all timing devices in my ship have to run slow in precisely the same way so that I cannot detect the motion of the ship."

The astronaut's observation highlights the power and generality of the principle of relativity. It turns a limited theory about the behavior of one special kind of clock into a general theory about the behavior of all possible clocks. If the light pulse clock in the astronaut's capsule is running slow by a factor  $\sqrt{1 - v^2/c^2}$ , then all clocks must run slow by exactly the same factor so that the astronaut cannot detect his motion.

## **Real Clocks**

Our theory still has a severe limitation. We have to assume that the light pulse clock runs slow. But no one has yet built a light pulse clock. Thus our theory is still based on thought experiments and conjectures about the behavior of light. If we had just one real clock that ran slow by a factor  $\sqrt{1 - v^2/c^2}$ , then the principle of relativity would guarantee that all other clocks ran slow in precisely the same way. Then we would not need any conjectures about the behavior of light. The principle of relativity would do it all!

In 1905 when Einstein proposed the special theory of relativity, he did not have any examples of moving clocks that were observed to run slow. He had to rely on his intuition and the two postulates. It was not until the early 1930s, in studies of the behavior of an elementary particle called the *muon*, that experimental evidence was obtained showing that a real moving clock actually ran slow.

A muon at rest has a half life of 2.2 microseconds or 2,200 nanoseconds. That means that if we start with 1000 muons, 2.2 microseconds later about half will have decayed and only about 500 will be left. Wait another 2.2 microseconds and half of the remaining muons will decay and we will have only about 250 left, etc. If we wait 5 half lives, just over 10 microseconds, only one out of 32 of the original particles remain  $(1/2 \times 1/2 \times 1/2 \times 1/2 \times 1/2 = 1/32)$ .



#### Figure 19

In our muon clock, every time half of the muons inside decay, we replace them and move the hand on the face forward by 2.2 microseconds.

Muons are created when cosmic rays from outer space strike the upper atmosphere. Few cosmic rays make it down to the lower atmosphere, so that most muons are created in the upper atmosphere, several miles up. The interesting results, observed in the 1930s was that there were almost as many high energy muons striking the surface of the earth as there were several miles up. This indicated that most of the high energy muons seemed to be surviving the several mile trip down through the earth's atmosphere.

Suppose we have a muon traveling at almost the speed of light, almost 1 foot per nanosecond. To go a mile, 5280 feet, would take 5,280 nanoseconds or about 5 microseconds. Therefore a 2 mile trip takes at least 10 microseconds, which is 5 half lives. One would expect that in this 2 mile trip, only one out of every 32 muons that started the trip would survive. Yet the evidence was that most of the high energy muons, those traveling close to the speed of light, survived. How did they do this?

We can get an idea of why the muons survive when we realize that the muon half life can be used as a timing device for a clock. Imagine that we have a box with a dial on the front as shown in Figure (19). We set the hand to 0 and put 1000 muons in the box. We wait until half the muons decay, whereupon we advance the hand 2.2 microseconds, replace the decayed muons so that we again have 1000 muons, and then wait until half have decayed again. If we keep repeating this process the hand will advance one muon half life in each cycle. Here we have a clock based on the muon half life rather than the swings of a pendulum or the vibrations of a quartz crystal.

The fact that most high energy muons raining down through the atmosphere survive the trip means that their half life is in excess of 10 microseconds, much longer than the 2.2 microsecond half life of a muon at rest. A clock based on these moving muons would run much slower than a muon clock at rest. Thus the experimental observation that the muons survive the trip down through the atmosphere gives us our first example of a real clock that runs slow when moving. In the early 1960s, a motion picture was made that carefully studied the decay of muons in the trip down from the top of Mount Washington in New Hampshire to sea level (the sea level measurements were made in Cambridge, Massachusetts), a trip of about 6000 feet. Muons traveling at a speed of v = .994c were studied and from the number surviving the trip, it was determined that the muon half life was lengthened to about 20 microseconds, a factor of 9 times longer than the 2.2 microsecond lifetime of muons at rest. Since  $1/\sqrt{1 - v^2/c^2} = 9$  for v = .994c, a result we got back in Equation (16), we see that the moving picture provides an explicit example of a moving clock that runs slow by a factor  $\sqrt{1 - v^2/c^2}$ .

At the present time there are two ways to observe the slowing down of real clocks. One is to use elementary particles like the muon, whose lifetimes are lengthened significantly when the particle moves at nearly the speed of light. The second way is to use modern atomic clocks which are so accurate that one can detect the tiny slowing down that occurs when the clock rides on a commercial jet. We calculated that a clock traveling 500 miles per hour should lose one nanosecond every hour. This loss was detected to an accuracy of 1% when physicists at the University of Maryland in the early 1980s flew an atomic clock for 15 hours over Chesapeake Bay.

In more recent times atomic clocks have become so accurate that the slowing down of the clock has become a nuisance. When these clocks are moved from one location to another, they have to be corrected for the time that was lost due to their motion. For these clocks, even a one nanosecond error is too much.

Thus today the slowing down of moving clocks is no longer a hypothesis but a common observational fact. The slowing down by  $\sqrt{1 - v^2/c^2}$  has been seen both for clocks moving at the slow speeds of a commercial jet and the high speeds travelled by elementary particles. We now have real clocks that run slow by a factor  $\sqrt{1 - v^2/c^2}$  and no longer need to hypothesize about the behavior of light pulses. *All of our conjectures in this chapter hinge on the principle of relativity alone.* 



#### Figure 19a -- Muon Lifetime Movie

The lifetimes of 568 muons, traveling at a speed of .994c, were plotted as vertical lines. If the muon's clocks did not run slow, these lines would show how far the muons could travel before decaying. One can see that very few of the muons would survive the trip from the top of Mt. Washington to sea level. Yet the majority do survive.

#### Movie

To play the movie, click the cursor in the photo to the left. Use up or down arrows on the keyboard to raise or lower volume. Left and right arrows step one frame foreward or back and esc stops it. The movie is 36 minutes long. The Movie Time Dilation: An Experiment with Mu-Mesons is presented with the permission of Education Development Center Inc., Newton, Massachusetts.

# **Time Dilation**

If all moving clocks run slow, does *time itself* run slow for the moving observer? That raises the question of how we define time. If time is nothing more than what we measure by clocks, and all clocks run slow, we might as well say that time runs slow. And we can give this effect a name like time dilation, the word dilation referring to the stretching out of seconds in a moving clock.

But time is such a personal concept, it plays such a basic role in our lives, that it seems almost demeaning that time should be nothing more than what we measure by clocks. We have all had the experience that time runs slow when we are bored, and fast when we are busy. Time is associated with all aspects of our life, including death. Can such an important concept be abstracted to be nothing more than the results of a series of measurements?

Let us take the following point of view. Let physicists' time be that which is measured by clocks. Physicists' time is what runs slow for an object moving by. If your sense of time does not agree with physicists' time, think of that as a challenge. Try to devise some experiment to show that your sense of time is measurably different from physicists' time. If it is, you might be able to devise an experiment that violates the principle of relativity.

# Space Travel

In human terms, time dilation should have its greatest effect on space travelers who need to travel long distances and therefore must go at high speeds. To get an idea of the distances involved in space travel, we note that light takes 1.25 seconds to travel from the earth to the moon (the moon is 1.25 billion feet away), and 8 minutes to travel from the sun to the earth. We can say that the moon is 1.25 light seconds away and the sun is 8 light minutes distant.

Currently Neptune is the most distant planet (Pluto will be the most distant again in a few years). When Voyager II passed Neptune, the television signals from Voyager, which travel at the speed of light, took 2.5 hours to reach us. Thus our solar system has a radius of 2.5 light hours. It takes 4 years for light to reach us from the nearest star from our sun; stars are typically one to a few light years apart.

If you look up at the sky at night and can see the Milky Way, you will see part of our galaxy, a spiral structure of stars that looks much like our neighboring galaxy Andromeda shown in Figure (20). Galaxies are about 100,000 light years across, and typically spaced about a



**Figure 20** The Andromeda galaxy, about a million light years away, and about 1/10 million light years in diameter.

million light years apart. As we will see there are even larger structures in space; there are interesting things to study on an even grander scale.

Could anyone who is reading this text survive a trip to explore our neighboring galaxy Andromeda, or just survive a trip to some neighboring star, say, only 200 light years away?

Before Einstein's theory, one would guess that the best way to get to a distant star would be to go so fast that the trip would not take very long. But now we have a problem. In Einstein's theory, the speed of light is a special speed. If we had the astronaut carry our light pulse clock at a speed greater than the speed of light, the light pulse could not remain in the clock. The astronaut would also notice that he could not keep the light pulse in the clock, and could use that fact to detect his own uniform motion. In other words, the principle of relativity implies that we or the astronaut cannot travel faster than the speed of light.

That the speed of light is a limiting speed is common knowledge to physicists working with elementary particles. Small particle accelerators about a meter in diameter can accelerate electrons up to speeds approaching v = .9999c. The two mile-long accelerator at Stanford University, which holds the speed record for accelerating elementary particles here on earth, can only get electrons up to a speed v = .9999999999c. The speed of light is Nature's speed limit, how this speed limit is enforced is discussed in Chapter 6. Does Einstein's theory preclude the possibility that we could visit a distant world in our lifetime; are we confined to our local neighborhood of stars by Nature's speed limit? The behavior of the muons raining down through the atmosphere suggests that we are not confined. The muons, you will recall, live only 2.2 microseconds (on the average) when at rest. Yet the muons go much farther than the 2200 feet that light could travel in a muon half life. They survive the trip down through the atmosphere because their clocks are running slow.

If humans could accompany muons on a trip at a speed v = .994c, the human clocks should also run slow, their lifetimes should also expand by the same factor of 9. If the human clocks did not run slow and the muon clocks did, the difference in rates could be used to detect uniform motion in violation of the principle of relativity.

The survival of the muons suggest that we should be able to travel to a distant star in our own lifetime. Suppose, for example, we wish to travel to the star Zeta (we made up that name) which is 200 light years away. If we traveled at the speed v = .994c, our clocks should run slow by a factor  $1/\sqrt{1 - v^2/c^2} = 1/9$ , and the trip should only take us  $200 \times 1/9 = 22.4$  years. We would be only 22.4 years older when we get there. A healthy, young crew should be able to survive that.

# **The Lorentz Contraction**

A careful study of this proposed trip to star Zeta uncovers a consequence of Einstein's theory that we have not discussed so far. To see what this effect is, to see that it is just as real as the slowing down of moving clocks, we will treat this proposed trip as a new thought experiment which will be analyzed from several points of view.

In this thought experiment, the instructor and the class, who participated in the previous thought experiments, decide to travel to Zeta at a speed of v = .994c. They have a space ship constructed which on the inside looks just like their classroom, so that classroom discussions can be continued during the trip.

On the earth, a permanent government subagency of NASA is established to record transmissions from the space capsule and maintain an earth bound log of the trip. Since the capsule, traveling at less than the speed of light, will take over 200 years to get to Zeta, and since the transmissions upon arrival will take 200 years to get back, the NASA agency has to remain in operation for over 400 years to complete its assignment. NASA's summary of the trip, written in the year 2406, reads as follows: "The spacecraft took off in the year 2001 and spent four years accelerating up to a speed of v = .994c. During this acceleration everything was packed away, but when they got up to the desired speed, the rocket engines were shut off and they started the long coast to the Zeta. This coast started with a close fly-by of the earth in late January of the year 2005. The NASA mission control officer who recorded the fly-by noted that his great, great, great, grandchildren would be alive when the spacecraft reached its destination."

The mission control officer then wrote down the following calculations that were later verified in detail. "The spacecraft is traveling at a speed v = .994c, so that it will take 1/.994 times longer than it takes a pulse of light to reach the star. Since the star is 200 light years away, the spacecraft should take 200/.994 = 201.2 years to get there.

But the passengers inside are also moving at a speed v = .994c, their clocks and biological processes run slow by a factor  $\sqrt{1 - v^2/c^2} = 1/9$ , and the amount of time they will age is

amount of time space travelers = 201.2 years  $\times \frac{1}{9}$  = 22.4 years age

Even the oldest member of the crew, the instructor, will be able to survive."

The 2406 entry continued; "During the intervening years we maintained communication with the capsule and everything seemed to go well. There were some complaints about our interpretation of what was happening but that did not matter, everything worked out just as we had predicted. The spacecraft flew past Zeta in March of the year 2206, and we received the communications of the arrival this past March. The instructor said she planned to retire after they decelerated and the spacecraft landed on a planet orbiting Zeta. She was not quite sure what her class of middle aged students would do."

NASA's predictions may have come true, but from the point of view of the class in the capsule, not everything worked out the way NASA said it did.

As NASA mentioned, a few years were spent accelerating the space capsule to the speed v = .994c. The orbit was chosen so that just after the engines were shut off and the coast to Zeta began, the spacecraft would pass close to the earth for one final good-by.

There was quite a change from the acceleration phase to the coasting phase. During the acceleration everything had to be securely fastened, and there was the constant vibration of the engines. But when the engines were shut off, you couldn't feel motion any more; everything floated as in the TV pictures of the early astronauts orbiting the earth.

When the coasting started, the instructor and class settled down to the business of monitoring the trip. The first step was to test the principle of relativity. Was there any experiment that they could do inside the capsule that could detect the motion of the capsule? Various experiments were tried, but none demonstrated that the capsule itself was moving. As a result the students voted to take the point of view that they and the capsule were at rest, and the things outside were moving by.



#### Figure 21

To measure the speed of the earth as it passes by, the class measures the time it takes a small satellite to pass by the windows in the back of the classroom. The windows are 100 feet apart. Very shortly after the engines were shut off, the earth went by. This was expected, and the students were ready to measure the speed of the earth as it passed. There were two windows 100 feet apart on the back wall of the classroom, as shown in Figure (21). When the earth came by, there was an orbiting spacecraft, essentially at rest relative to the earth, that passed close to the windows. The students measured the time it took the front edge of this orbiting craft to travel the 100 feet between the windows. They got 100.6 nanoseconds and therefore concluded that the orbiting craft and the earth itself were moving by at a speed

$$v_{earth} = \frac{100 \text{ feet}}{100.6 \text{ nanoseconds}}$$
  
= .994  $\frac{\text{feet}}{\text{nanosecond}}$  = .994 c

So far so good. That was supposed to be the relative speed of the earth.

In the first communications with earth, NASA mission control said that the space capsule passed by the earth at noon, January 17, 2005. Since all the accurate clocks had been dismantled to protect them from the acceleration, and only put back together when the coasting started, the class was not positive about what time it was. They were willing to accept NASA's statement that the fly-by occurred on January 17, 2005. From then on, however, the class had their own clocks in order—light pulse clocks, digital clocks and an atomic clock. From then on they would keep their own time.

For the next 22 years the trip went smoothly. There were numerous activities, video movies, etc., to keep the class occupied. Occasionally, about once every other month, a star went by. As each star passed, its speed v was measured and the class always got the answer v = .994c. This confirmed that the earth and the neighboring stars were all moving together like bright dots on a huge moving wall.

The big day was June 13, 2027, the 45th birthday of Jill who was eighteen when the trip was planned. This was the day, 22.4 years after the earth fly-by, that Zeta went by. The students made one more speed measurement and determined that Zeta went by at a speed v = .994c. An arrival message was sent to NASA, one day was allowed for summary discussions of the trip, and then the deceleration was begun.

After a toast to Jill for her birthday, Bill began the conversation. "Over the past few years, the NASA communications and even our original plans for the trip have been bothering me. The star charts say that Zeta is 200 light years from the earth, but that cannot be true."

"Look at the problem this way." Bill continues. "The earth went by us at noon on January 17, 2005, just 22.4 years ago. When the earth went by, we observed that it took 100.6 nanoseconds to pass by our 100 foot wide classroom. Thus the earth went by at a speed v = .994 feet/nanosecond, or .994c. Where is the earth now, 22.4 years later? How far could the earth have gotten, traveling at a speed .994c for 22.4 years? My answer is

distance of earth from spaceship =  $.994 \frac{\text{light year}}{\text{year}} \times 22.4 \text{ years}$ 

= 22 light years

"You're right!" Joan interrupted, "Even if the earth had gone by at the speed of light, it would have gone only 22.4 light years in the 22.4 years since fly-by. The star chart must be wrong."

The instructor, who had just entered the room, said, "I object to that remark. As a graduate student I sat in on part of a course in astronomy and they described how the distance to Zeta was measured." The instructor drew a sketch, Figure (22), and continued. "Here is the earth in its orbit about the sun, and two observations, six months apart, are made of Zeta. You see that the two positions of the earth and the star form a triangle. Telescopes can accurately measure the two angles I labeled  $\theta_1$  and  $\theta_2$ , and the distance across the earth's orbit is accurately known to be 16 light minutes. If you know two angles and one side of a triangle, then you can calculate the other sides from simple geometry. One reason for choosing a trip to Zeta is that we had accurate

measurements of the distance to that star. We knew that it was 200 light years away, and we knew that traveling at a speed .994c, we could survive the trip in our lifetime."

Bill responded, "I think you entered the room too late and missed my argument. Let me summarize it. *Point 1*: the earth went by a little over 22 years ago. *Point* 2: we actually measured that the earth was traveling by us at almost the speed of light. *Point* 3: even light cannot go farther than 22 light years in 22 years. The earth can be no farther than about 22 light years away. *Point* 4: Zeta passed by us today, thus *the distance from the earth to Zeta is about 22 light years, not 200 light years!*"

"But what about NASA's calculations and all their plans," the instructor said, interrupting a bit nervously.

"We do not care what NASA thinks," responded Bill. "We have had no acceleration since the earth went by. Thus the principle of relativity guarantees that we can take the point of view that we are at rest and that it is the earth and NASA that are moving. From our point of view, the earth is 22 light years away. What NASA thinks is their business."

Joan interrupts, "Let us not argue on this last day. Let's figure out what is happening. There is something more important here than just how far away the earth is."



Figure 22

Instructor's sketch showing how the distance from the earth to the star Zeta was measured. (For a star 200 light years away,  $\theta_3$  is 4.5 millionths of a degree.)
"Remember in the old lectures on time dilation where the astronaut carried a light pulse clock. We used the peculiar behavior of that clock and the principle of relativity to deduce that time ran slow for a moving observer."

"Now for us, NASA is the moving observer. More than that, the earth, sun, and the stars, including Zeta, have all passed us going in the same direction and the same speed v = .994c. We can think of them as all in the same huge space ship. Or we can think of the earth and the stars as painted dots on a very long rod. A very long rod moving past us at a speed v = .994c. See my sketch (Figure 23)."

"To NASA, and the people on earth, this huge rod, with the sun at one end and Zeta at the other, is 200 light years long. Our instructor showed us how earth people measured the length of the rod. But as Bill has pointed out, to us this huge rod is only 22 light years long. That moving rod is only 1/9th as long as the earth people think it is."

"But," Bill interrupts, the factor of 1/9 is exactly the factor  $\sqrt{1 - v^2/c^2}$  by which the earth people thought our clocks were running slow. Everyone sees something peculiar. The earth people see our clocks running slow by a factor  $\sqrt{1 - v^2/c^2}$ , and we see this hypothetical rod stretched from the sun to Zeta contracted by a factor  $\sqrt{1 - v^2/c^2}$ ."

"But I still worry about the peculiar rod of Joan's," Bill continues, "what about real rods, meter sticks, and so forth? Will they also contract?"

At this point Joan sees the answer to that. "Remember, Bill, when we first discussed moving clocks, we had only the very peculiar light pulse clock that ran slow. But then we could argue that all clocks, no matter how they are constructed, had to run slow in exactly the same way, or we could violate the principle of relativity."

"We have just seen that my 'peculiar' rod, as you call it, contracts by a factor  $\sqrt{1 - v^2/c^2}$ . We should be able to show with some thought experiments that all rods, no matter what they are made of, must contract in exactly the same way as my peculiar one or we could violate the principle of relativity." "That's easy," replies Bill. Just imagine that we string high tensile carbon filament meter sticks between the sun and Zeta. I estimate (after a short calculation) that it should take only  $6 \times 10^{17}$  of them. As we go on our trip, it doesn't make any difference whether the meter sticks are there or not, everything between the earth and Zeta passes by in 22 years. We still see  $6 \times 10^{17}$  meter sticks. But each one must have shortened by a factor  $\sqrt{1 - v^2/c^2}$  so that all of them fit in the shortened distance of 22 light years. It does not make any difference what the sticks are made of."

Jim, who had not said much up until now, said, "OK, from your arguments I see that the length of the meter sticks, the length in the direction of motion must contract by a factor  $\sqrt{1 - v^2/c^2}$ , but what about the width? Do the meter sticks get skinnier too?"

The class decided that Jim's question was an excellent one, and that a new thought experiment was needed to decide.

Let's try this," suggested Joan. "Imagine that we have a space ship 10 feet in diameter and we build a brick wall with a circular hole in it 10 feet in diameter (Figure (24)). Let us assume that widths, as well as lengths, contract. To test the hypothesis, we hire an astronaut to



#### Joan's sketch of the Sun and Zeta moving by. This "object" passed by in about 22 years, moving at nearly the speed of light. Thus the "object" was about 22 light years long.

fly the 10 foot diameter capsule through the 10 foot hole at nearly the speed of light, say at v = .994c. If widths contract like lengths, the capsule should contract to 10/ 9 of a foot; it should be just over 13 inches in diameter when it gets to the 10 foot hole. It should have no trouble getting through."

"But look at the situation from the astronaut's point of view. He is sitting there at rest, and a brick wall is approaching him at a speed v = .994c. He has been told that there is a 10 foot hole in the wall, but he has also been told that the width of things contracts by a factor  $\sqrt{1 - v^2/c^2}$ . That means that the diameter of the hole should contract from 10 feet to 13 inches. He is sitting there in a 10 foot diameter capsule, a brick wall with a 13 inch hole is approaching him, and he is supposed to fit through. No way! He bails out and looks for another job."

"That's a good way to do thought experiments, Joan," replied the instructor. "Assume that what you want to test is correct, and then see if you can come up with an inconsistency. In this case, by assuming that widths contract, you predicted that the astronaut should easily make it through the hole in the wall. But the astronaut faced disaster. The crash, from the astronaut's point of view would have been an unfortunate violation of the principle of relativity, which he could use as evidence of his own uniform motion."

"To sum it up," the instructor added, "we now have time dilation where moving clocks run slow by a factor  $\sqrt{1 - v^2/c^2}$ , and we see that moving lengths contract by the same factor. Only lengths in the direction of motion contract, widths are unchanged."



**Figure 24** *Do diameters contract?* 

Leaving our thought experiment, it is interesting to note that the discovery of the contraction of moving lengths occurred before Einstein put forth the special theory of relativity. In the 1890s, physicist George Fitzgerald assumed that the length of one of the arms in Michaelson's interferometer, the arm along the direction of motion, contracted by a factor  $\sqrt{1 - v^2/c^2}$ . This was just the factor needed to keep the interferometer from detecting the earth's motion in the Michaelson-Morley experiments. It was a short while later that H.A. Lorentz showed that if the atoms in the arm of the Michaelson interferometer were held together by electric forces, then such a contraction would follow from Maxwell's theory of electricity. The big step, however, was Einstein's assumption that the principle of relativity is correct. Then, if one object happens to contract when moving, all objects must contract in exactly the same way so that the contraction could not be used to detect one's own motion. This contraction is called the Lorentz-Fitzgerald contraction, or Lorentz contraction. for short.

## **Relativistic Calculations**

Although we have not quite finished with our discussion of Einstein's special theory of relativity, we have covered two of the important consequences, time dilation and the Lorentz contraction, which will play important roles throughout the text. At this point we will take a short break to discuss easy ways to handle calculations involving these relativistic effects. Then we will take another look at Einstein's theory to see if there are any more new effects to be discovered.

After our discussion of time dilation, we pointed out the importance of the quantity  $\sqrt{1 - v^2/c^2}$  which is a number always less than 1. If we wanted to know how much longer a moving observer's time interval was, we divided by  $\sqrt{1 - v^2/c^2}$  to get a bigger number. If we wanted to know how much less was the frequency of a moving clock, we multiplied by  $\sqrt{1 - v^2/c^2}$  to get a smaller number.

With the Lorentz contraction we have another effect that depends upon  $\sqrt{1 - v^2/c^2}$ . If we see an object go by us, the object will contract in length. To predict its contracted length, we multiply the uncontracted length by  $\sqrt{1 - v^2/c^2}$  to get a smaller number. If, on the other

hand, an object moving by us had a contracted length l, and we stop the object, the contraction is undone and the length increases. We get the bigger uncontracted length by dividing by  $\sqrt{1 - v^2/c^2}$ .

As we mentioned earlier, first determine intuitively whether the number gets bigger or smaller, then either multiply by or divide by the  $\sqrt{1 - v^2/c^2}$  as appropriate. This always works for time dilation, the Lorentz contraction, and, as we shall see later, relativistic mass.

We will now work some examples involving the Lorentz contraction to become familiar with how to handle this effect.

#### **Example 1** Muons and Mt Washington

In the Mt. Washington experiment, muons travel 6000 feet from the top of Mt. Washington to sea level at a speed v = .994c. Most of the muons survive despite the fact that the trip should take about 6 microseconds (6000 nanoseconds), and the muon half life is  $\tau = 2.2$  microseconds for muons at rest.

We say that the muons survive the trip because their internal timing device runs slow and their half life expands by a factor  $1/\sqrt{1 - v^2/c^2} = 9$ . The half life of the moving muons should be

half life of  
moving muons = 
$$\frac{\tau}{\sqrt{1 - v^2/c^2}}$$
  
= 2.2 microseconds × 9  
= 19.8 microseconds

This is plenty of time for the muons to make the trip.

From the muon's point of view, they are sitting at rest and it is Mt. Washington that is going by at a speed v =.994c. The muon's clocks aren't running slow, instead the height of Mt. Washington is contracted.

To calculate the contracted length of the mountain, start with the 6000 foot uncontracted length, multiply by  $\sqrt{1 - v^2/c^2} = 1/9$  to get

contracted height  
of Mt. Washington = 
$$6000 \text{ feet} \times \frac{1}{9}$$
  
=  $667 \text{ feet}$ 

Traveling by at nearly the speed of light, the 667 foot high Mt. Washington should take about 667 nanoseconds or .667 microseconds to go by. Since this is considerably less than the 2.2 microsecond half life of the muons, most of them should survive until sea level comes by.

#### Example 2 Slow Speeds

Joan walks by us slowly, carrying a meter stick pointing in the direction of her motion. If her speed is v = 1 foot/second, what is the contracted length of her meter stick as we see it?

This is an easy problem to set up. Since her meter stick is contracted, we multiply 1 meter times the  $\sqrt{1 - v^2/c^2}$ with v = 1 foot/second. The problem comes in evaluating the numbers. Noting that 1 nanosecond = 10<sup>-9</sup> seconds, we can use the conversion factor 10<sup>-9</sup> seconds/nanosecond to write

$$v = 1 \frac{\text{ft}}{\text{sec}} \times 10^{-9} \frac{\text{sec}}{\text{nanosecond}}$$
$$= 10^{-9} \frac{\text{ft}}{\text{nanosecond}} = 10^{-9} \text{ c}$$

Thus we have

$$\frac{v}{c} = 10^{-9}, \frac{v^2}{c^2} = 10^{-18}$$

and for Joan's slow walk we have

$$\sqrt{1 - v^2/c^2} = \sqrt{1 - 10^{-18}} \tag{19}$$

If we try to use a calculator to evaluate the square root in Equation (19), we get the answer 1. For the calculator, the number  $10^{-18}$  is so small compared to 1, that it is ignored. It is as if the calculator is telling us that when Joan's meter stick is moving by at only 1 foot/second, there is no noticeable contraction.

But there is some contraction, and we may want to know the contraction no matter how small it is. Since calculators cannot handle numbers like  $1 - 10^{-18}$ , we need some other way to deal with such expressions. For this, there is a convenient set of approximation formulas which we will now derive.

## **Approximation Formulas**

The approximation formulas deal with numbers close to 1, numbers that can be written in the form (1 + a) or (1 - a) where a is a number much less than 1. For example the square root in Equation (19) can be written as

$$\sqrt{1 - 10^{-18}} = \sqrt{1 - \alpha}$$

where  $\alpha = 10^{-18}$  is truly a number much less than 1.

The idea behind the approximation formulas is that if a is much less than 1,  $a^2$  is very much less than 1 and can be neglected. To see how this works, let us calculate (1 + a)<sup>2</sup> and see how we can neglect  $a^2$  terms even when a is as large as .01. An exact calculation is

 $(1 + \alpha)^2 = 1 + 2\alpha + \alpha^2$ 

which for  $\alpha = .01$ ,  $\alpha^2 = .0001$  is

$$(1 + \alpha)^2 = 1 + .02 + .0001$$
  
= 1.0201

If we want to know how much  $(1 + \alpha)^2$  differs from 1, but do not need too much precision, we could round off 1.0201 to 1.02 to get

$$(1+\alpha)^2 \approx 1.02$$

(The symbol  $\approx$  means "approximately equal to"). But in replacing 1.0201 by 1.02, we are simply dropping the  $\alpha^2$  term in Equation (19). We can write

$$(1 + \alpha)^2 \approx 1 + 2\alpha = 1 + .02 = 1.02$$
 (20)

Equation (20) is our first example of an approximation formula.

In Equation (20) the smaller a is the better the approximation. If a = .0001 we have

$$(1.0001)^2 = 1.00020001$$
 (exact)

Equation (20) gives

$$(1 + .0001)^2 \approx 1 + .0002 = 1.0002$$

and we see that the neglected  $\alpha^2$  terms become less and less important.

Some useful approximation formulas are the following

$$(1+\alpha)^2 \approx 1+2\alpha \tag{20}$$

$$(1 - \alpha)^2 \approx 1 - 2\alpha$$
 (21)

$$\frac{1}{1+\alpha} \approx 1-\alpha \tag{22}$$

$$\frac{1}{1-\alpha} \approx 1+\alpha \tag{23}$$

$$\sqrt{1 - \alpha} \approx 1 - \frac{\alpha}{2}$$
 (24)

$$\frac{1}{\sqrt{1-\alpha}} \approx 1 + \frac{\alpha}{2}$$
(25)

We have already derived Equation (20). Equation (21) follows from (20) if we replace  $\alpha$  by  $-\alpha$ .

Equation (22) can be derived as follows. Multiply the quantity 1 - a by (1+a)/(1+a) which is 1 to get

$$1 - \alpha = (1 - \alpha) \times \frac{(1 + \alpha)}{(1 + \alpha)} = \frac{(1 - \alpha^2)}{1 + \alpha} \approx \frac{1}{1 + \alpha}$$

In the last step we dropped the  $\alpha^2$  terms.

To derive the approximate formula for a square root, start with

$$\left(1-\frac{\alpha}{2}\right)\times\left(1-\frac{\alpha}{2}\right) = 1-2\frac{\alpha}{2}+\frac{\alpha^2}{4} \approx 1-\alpha$$
 (26)

taking the square root of Equation (26) gives

$$1 - \frac{\alpha}{2} \approx \sqrt{1 - \alpha}$$

which is the desired result. Again we only neglected  $\alpha^2$  terms.

To derive Equation (25), first use Equation (24) to get

$$\frac{1}{\sqrt{1-\alpha}} \approx \frac{1}{1-\frac{\alpha}{2}}$$

Then use Equation (23), with a replaced by a/2 to get

$$\frac{1}{1-\frac{\alpha}{2}} \approx 1 + \frac{\alpha}{2}$$

which is the desired result.

For those who are interested, the approximation formulas we have written are the first term of the so called *binomial expansion*:

$$(1 + \alpha)^n = 1 + n\alpha + \frac{n(n-1)}{2}\alpha^2 + \cdots$$
 (27)

where the coefficients of  $\alpha$ ,  $\alpha^2$ , etc. are known as the *binomial coefficients*. If you need more accurate approximations, you can use Equation (27) and keep terms in  $\alpha^2$ ,  $\alpha^3$ , etc. For all the work in this text, the first term is adequate.

#### **Exercise 5**

Show that Equations (20) through (25) are all examples of the first order binomial expansion

$$(1+\alpha)^{n} \approx 1 + n\alpha \tag{27a}$$

We are now ready to apply our approximation formulas to evaluate  $\sqrt{1 - 10^{-18}}$  that appeared in Equation (17). Since  $\alpha = 10^{-18}$  is very small compared to 1, we have

$$\sqrt{1 - 10^{-18}} = \sqrt{1 - \alpha} \approx 1 - \frac{\alpha}{2} = 1 - \frac{10^{-18}}{2}$$

Thus the length of Joan's meter stick is

length of Joan's contracted meter stick = 1 meter  $\times \sqrt{1 - v^2/c^2}$ 

$$= 1 \operatorname{meter}\left(1 - \frac{10^{-18}}{2}\right)$$

= 1 meter -  $5 \times 10^{-19}$  meters

#### **Exercise 6**

We saw that time dilation in a commercial jet was not a big effect either—clocks losing only one nanosecond per hour in a jet traveling at 500 miles per hour. This was not an unnoticed effect, however, because modern atomic clocks can detect this loss.

In our derivation of the one nanosecond loss, we stated in Equation (17) that

$$\frac{1}{\sqrt{1 - v^2/c^2}} \approx 1 + 2.7 \times 10^{-13} \qquad \begin{array}{c} \text{for a speed of} \\ 500 \text{ miles/hour} \end{array}$$
(17)

Starting with

$$v = 500 \frac{\text{miles}}{\text{hour}} \times 5280 \frac{\text{feet}}{\text{mile}} \times \frac{1}{3600 \text{ sec/hour}}$$

use the approximation formulas to derive the result stated in Equation (17).

#### **Exercise 7**

Here is an exercise where you do not need the approximation formulas, but which should get you thinking about the Lorentz contraction. Suppose you observe that the Mars-17 spacecraft, traveling by you at a speed of v = .995c, passes you in 20 nanoseconds. Back on earth, the Mars-17 spacecraft is stored horizontally in a hanger that is the same length as the spacecraft. How long is the hanger?

## **A CONSISTENT THEORY**

As we gain experience with Einstein's special theory of relativity, we begin to see a consistent pattern emerge. We are beginning to see that there is general agreement on what happens, even if different observers have different opinions as to *how* it happens. A good example is the Mt. Washington experiment observing muons traveling from the top of Mt. Washington to sea level. Everyone agrees that the muons made it. The muons are actually seen down at sea level. How they made it is where we get the differing points of view. We say that they made it because their clocks ran slow. They say they made it because the mountain was short. Time dilation is used from one point of view, the Lorentz contraction from another.

Do we have a complete, consistent theory now? In any new situation will we always agree on the predicted outcome of an experiment, even if the explanations of the outcome differ? Or are there some new effects, in addition to time dilation and the Lorentz contraction, that we will have to take into account?

The answer is that there is one more effect, called the *lack of simultaneity* which is a consequence of Einstein's theory. When we take into account this lack of simultaneity as well as time dilation and the Lorentz contraction, we get a completely consistent theory. Everyone will agree on the predicted outcome of any experiment involving uniform motion. No other new effects are needed to explain inconsistencies.

The lack of simultaneity turns out to be the biggest effect of special relativity, it involves two factors of  $\sqrt{1 - v^2/c^2}$ . But in this case the formulas are not as important as becoming familiar with some of the striking consequences. We will find ourselves dealing with problems such as whether we can get answers to questions that have not yet been asked, or whether gravity can crush matter out of existence. Strangely enough, these problems are related.

## LACK OF SIMULTANEITY

One of the foundations of our intuitive sense of time is the concept of simultaneity. "Where were you when the murder was committed," the prosecutor asks. "At the time of the murder," the defendant replies, "I was eating dinner across town at Harvey's Restaurant." If the defendant can prove that the murder and eating dinner at Harvey's were simultaneous events, the jury will set him free. Everyone knows what simultaneous events are, or do they?

One of the most unsettling consequences of Einstein's theory is that the simultaneity of two events depends upon the point of view of the observer. Two events that from our point of view occurred simultaneously, may not be simultaneous to an observer moving by. Worse yet, two events that occurred one after the other to us, may have occurred in the reverse order to a moving observer.

To see what happens to the concept of simultaneous events, we will return to our thought experiment involving the instructor and the class. The action takes place on the earth before the trip to the star Zeta, and Joan has just brought in a paperback book on relativity.

"I couldn't understand that book either," the instructor says to Joan, "he starts with Einstein's analogy of trains and lightening bolts, but then switches to wind and sound waves, which completely confused me. There are many popular attempts to explain Einstein's theory, but most do not do very well when it comes to the lack of simultaneity."

"One of the problems with these popular accounts," the instructor continues, "is that we have to imagine too much. In today's lecture I will try to avoid that. In class we are going to carry out a real experiment involving two simultaneous events. We are going to discuss that experiment until everyone in class is completely clear about what happened. No imagining yet, just observe what actually occurred. When there are no questions left, then we will look at our real experiment from the point of view of someone moving by. At that point the main features of Einstein's theory are easy to see." "The apparatus for our experiment is set up here on the lecture bench (Figure 25). On the left side of the bench I have a red flash bulb and on the right side a green flash bulb. These flash bulbs are attached to batteries and photocells so that when a light beam strikes their base, they go off."

"In the center of the desk is a laser and in front of it a beam splitter that uses half silvered mirrors. When I turn the laser on, the laser beam comes out, strikes the beam splitter, and divides into two beams. One beam travels to the left and sets off the red flash bulb, while the other beam goes to the right and triggers the green flash bulb. I will call the beams emerging from the beam splitter 'trigger beams' or 'trigger pulses'."

"Let us analyze the experiment before we carry it out," the instructor continues. "We will use the Einstein postulate that the speed of light is c to all observers. Thus the left trigger pulse travels at a speed c and so does the right one as I showed on the sketch. Since the beam splitter is in the center of the desk, the trigger pulses which start out together, travel the same distance at the same speeds to reach the flash bulbs. As a result the flash bulbs must go off simultaneously."

"The flashing of the flashbulbs are an example of what I mean by *simultaneous events*," the instructor adds with emphasis. "I know that they will be simultaneous events because of the way I set up the experiment".



#### Figure 25

Lecture demonstration experiment in which two flashbulbs are fired simultaneously by trigger signals from a laser. The laser and beam splitter are at the center of the lecture bench, so that the laser light travels equal distances to reach the red and green bulbs. A photocell, battery and relay are mounted in each flashbulb base. "OK, let's do the experiment."

While the instructor is adjusting the apparatus, one of the flashbulbs goes off accidentally which amuses the class, but finally the apparatus is ready, the laser beam turned on, and both bulbs fire.

"Well, were they simultaneous flashes?" the instructor asks the class.

"I guess so," Bill responds, a bit hesitantly.

"How do you know," the instructor asks.

"Because you set it up that way," answers Bill.

Turning and pointing a finger at Joan who is sitting on the right side of the room nearer the green flash bulb (as in Figure 26), the instructor says, "Joan, for you which flash was first?" Joan thought for a second and replied, "The green bulb is closer, I should have seen the green light first."

"But which occurred first?" the instructor interrupts.

"What are you trying to get at?" Joan asks.



#### Figure 26

Although Joan sees the light from the green flash first, she knows that the two flashes were simultaneous because of the way the experiment was set up.



"Let me put it this way," the instructor responds. "Around 1000 BC, the city of Troy fell to the invading Greek army. About the same time, a star at the center of the Crab Nebula exploded in what is known as a supernova explosion. Since the star is 2000 light years away, the light from the supernova explosion took 2000 years to get here. The light arrived on July 4, 1057, about the time of the Battle of Hastings. Now which are simultaneous events? The supernova explosion and the Battle of Hastings, or the supernova explosion and the fall of Troy"

"I get the point," replied Joan. "Just because they saw the light from the supernova explosion at the time of the Battle of Hastings, does not mean that the supernova explosion and that battle occurred at the same time. We have to calculate back and figure out that the supernova explosion occurred about the time the Greeks were attacking Troy, 2000 years before the light reached us."

"As I sit here looking at your experiment," Joan continues, "I see the light from the green flash before the light from the red flash, but I am closer to the green bulb than the red bulb. If I measure how much sooner the green light arrives, then measure the distances to the two bulbs, and do some calculations, I'll probably find that the two flashes occurred at the same time."



#### Figure 27a

In our thought experiment, a Martian astronaut passes by our lecture bench at a high speed v.

"It is much easier than that." the instructor exclaimed, "Don't worry about when the light reaches you, just look at the way I set up the experiment – two trigger pulses, starting at the same time, traveling the same distance at the same speed. The flashes must have occurred simultaneously. I chose this experiment because it is so easy to analyze when you look at the trigger pulses."

"Any other questions?" the instructor asks. But by this time the class is ready to go on. "Now let us look at the experiment from the point of view of a Martian moving to the right a high speed v (Figure 27a). The Martian sees the lecture bench, laser, beam splitter and two flash bulbs all moving to the left as shown (Figure 27b). The lecture bench appears shortened by the Lorentz contraction, but the beam splitter is still in the middle of the bench. What is important is that the trigger pulses, being light, both travel outward from the beam splitter *at a speed c*.

As the bench passes by, the Martian sees that the green flash bulb quickly runs into the trigger pulse like this (- c > v - c). But on the other side there is a race between the trigger pulse and the red flash bulb, (v - c), and the race continues for a long time after the green bulb has fired. For the Martian, the green bulb actually fired first, and the two flashes were not simultaneous."



#### Figure 27b

The Martian astronaut sees the green flashbulb running into its trigger signal and firing quickly. The red flashbulb is running away from it's trigger signal, and therefore will not fire for a long time. Clearly the green flash occurs first. "How much later can the red flash occur?" asks Bill.

The instructor replied, "The faster the bench goes by, the closer the race, and the longer it takes the trigger pulse to catch the red flash bulb. It isn't too hard to calculate the time difference. In the notes I handed out before class, I calculated that if the Martian sees our 12 foot long lecture bench go by at a speed

v = .9999999999999999999999999999992c (28)

then the Martian will determine that the red flash occurred one complete earth year after the green flash. Not only are the two flashes not simultaneous, there is no fundamental limit as to how far apart in time that the two flashes can occur."

The reader will find the instructor's class notes in Appendix A of this chapter.

At this point Joan asks a question. "Suppose an astronaut from the planet Venus passed our experiment traveling the other way. Wouldn't she see the red flash first?" "Let's draw a sketch," the instructor replies. The result is in Figure (28b). "The Venetian astronaut sees the lecture bench moving to the left. Now the red flash bulb runs into the trigger signal, and the race is with the green flash bulb. If the Venetian were going by at the same speed as the Martian (Equation 30) then the green flash would occur one year after the red one."

"With Einstein's theory, not only does the simultaneity of two events depend upon the observer's point of view, *even the order of the two events*—which one occurred first—depends upon how the observer is moving!"



#### Figure 28a

Now a Venusian astronaut passes by our lecture bench at a high speed v in the other direction.



#### Figure 28b

The Venetian astronaut sees the red flashbulb running into its trigger signal and firing quickly. The green flashbulb is running away from its trigger signal, and therefore will not fire for a long time. Clearly the red flash occurs first.

#### Figure 29

To test the speed of the computer, Bill thinks of a question, and types it in, when he sees the red flash. Joan checks to see if the answer arrives at the same time as the green flash.



## CAUSALITY

"You can reverse the order of two events that are years apart!" Bill exclaimed. "Couldn't something weird happen in that time?"

"What about cause and effect," asked Joan. "If you can reverse the order of events, can't you reverse cause and effect? Can't the effect come before the cause?"

"In physics," the instructor responds, "there is a principle called *causality* which says that you cannot reverse cause and effect. Causality is not equivalent to the principle of relativity, but it is closely related, as we can see from the following thought experiment."

"Suppose," she said, "we read an ad for a brand new IBM computer that is really fast. The machine is so fast that when you type a question in at one end, the answer is printed out at the other end, 4 nanoseconds later. We look at the ad, see that the machine is 12 feet long, and order one to replace our lecture bench. After the machine is installed, we decide to test the accuracy of the ad. Do we really get answers in 4 nanoseconds? To find out, we set up the laser, beam splitter and flash bulbs on the computer instead of the lecture bench. The

main difference in the setup is that the laser and beam splitter have been moved from the center, over closer to the end where we type in questions. We have set it up so that the trigger pulse travels 4 feet to the red bulb and 8 feet to the green bulb as shown in the sketch (Figure 29). Since the trigger pulse takes 4 nanoseconds to get to the red bulb, and 8 nanoseconds to reach the green bulb, the red flash will go off 4 nanoseconds before the green one. We will use these 4 nanoseconds to time the speed of the computer."

"Bill," the instructor says, motioning to him, "you come over here, and when you see the red flash *think* of a question. Then type it into the machine. Do not think of the question until after you see the red flash, but then think of it and type it in quickly. We will assume that you can do that in much less than a nanosecond. You can always do that kind of thing in a thought experiment."

"OK, Joan," the instructor says, motioning to Joan, "you come over here and look for the answer to Bill's question. If the ad is correct, if the machine is so fast that the answer comes out in 4 nanoseconds, then the answer should arrive when the green flash goes off."

#### Figure 30

To a Martian passing by, our computer is moving to the left at a speed near the speed of light. The race between the red bulb and its trigger signal takes so long that the green bulb fires first. As a result, Joan sees the answer to a question that Bill has not yet thought of. (This is what could happen if information travels faster than the speed of light.)



The instructor positions Bill and Joan and the equipment as shown in Figure (29), turns on the laser and fires the flash bulbs.

"Did you type in the question," the instructor asks Bill, "when the red flash occurred?"

"Of course," responds Bill, humoring the instructor.

"And did the answer arrive at the same time as the green flash," the instructor asks Joan.

"Sure," replies Joan, "why not?"

"Suppose it did," replied the instructor. "Suppose the ad is right, and the answer is printed when the green flash goes off. Let us now look at this situation from the point of view of a Martian who is traveling to the right at a very high speed. The situation to the Martian looks like this (Figure 30). Although the red bulb is closer to the beam splitter, it is racing away from the trigger pulse ( $v \leftarrow c \leftarrow c$ ). If the computer is going by fast enough, the race between the red bulb and its trigger pulse will take much longer than the head-on collision between the green bulb and its trigger pulse. The green flash will occur before the red flash."

"And I," interrupts Joan, "will see the answer to a question that Bill has not even thought of yet!"

"I thought you would be in real trouble if you could reverse the order of events," Joan added.

"It is not really so bad," the instructor continued. "If the ad is right, if the 12 foot long computer can produce answers that travel across the machine in 4 nanoseconds, we are in deep trouble. In that case we could see answers to questions that have not yet been asked. That machine can be used to violate the principle of causality. But there was something peculiar about that machine. When the answer went through the machine, information went through the machine at three times the speed of light. Light takes 12 nanoseconds to cross the machine, while the answer went through in 4 nanoseconds."

"Suppose," asks Bill, "that the answer did not travel faster than light. Suppose it took 12 nanoseconds instead of 4 nanoseconds for the answer to come out." The instructor replied, "To measure a 12 nanosecond delay with our flash bulb apparatus, we would have to set the beam splitter right up next to the red bulb like this (Figure 31) in order for the trigger signal to reach the green bulb 12 nanoseconds later. But with this setup, the red bulb flashes as soon as the laser is turned on. No one, no matter how they are moving by, sees a race between the red bulb and the trigger signal. Everybody agrees that the green flash occurs after the red flash."

# "You mean," interrupts Joan, "that you cannot violate causality if information does not travel faster than the speed of light?"

"That's right," the instructor replies, "that's one of the important and basic consequences of Einstein's theory."

"That's interesting," adds Bill. "It would violate the principle of relativity if we observed the astronaut's capsule, or probably any other object, traveling faster than the speed of light. The speed of light is beginning to play an important role."

#### Figure 31

If the answer to Bill's question takes 12 nanoseconds to travel through the 12 foot long computer, then this is the setup required to check the timing. The red bulb fires instantaneously, and everyone agrees that the red flash occurs first, and the answer appears later.



"That's pretty far out," replied Joan. "I didn't know that physics could say anything about how information – *ideas* – moved."

Jim, who had been sitting in the back of the classroom and not saying much, raised his hand. "At the beginning of the course when we were talking about sound pulses, you said that the more rigid the material, the faster the speed of sound in the material. You used Slinky pulses in your demonstrations because a Slinky is so compressible that a Slinky pulse travels slowly. You can't compress air as easily as a Slinky, and sound pulses travel faster in air. Since steel is very rigid, sound goes very fast in steel."

"During these discussions about the speed of light, I have been wondering. Is there any kind of material that is so rigid that sound waves travel at the speed of light?"

"What made you ask that?" the instructor asked.

"I've been reading a book about the life and death of stars," Jim replied. I just finished the chapter on neutron stars, and they said that the nuclear matter in a neutron star was very incompressible. It had to be to resist the strong gravitational forces. I was wondering, how fast is the speed of sound in this nuclear matter?"

"Up close to the speed of light," the instructor replied.

"If the nuclear matter were even more rigid, more incompressible, would the speed of sound exceed the speed of light?" Jim asked.

"It can't," the instructor replied.

"Then," Jim asked, "doesn't that put a limit on how incompressible, how rigid matter can be?"

"That looks like one of the consequences of Einstein's theory," the instructor replies.

"Then that explains what they were trying to say in the next chapter on black holes. They said that if you got too much matter concentrated in a small region, the gravitational force would become so great that it would crush the matter out of existence."

"I didn't believe it, because I thought that the matter would be squeezed down into a new form that is a lot more incompressible than nuclear matter, and the collapse of the star would stop. But now I am beginning to see that there may not be anything much more rigid than nuclear matter. Maybe black holes exist after all."

"Will you tell us about neutron stars and black holes?" Joan asks eagerly.

"Later in the course," the instructor responds.

## **APPENDIX A**

#### **Class Handout**

To predict how long it takes for the trigger pulse to catch the red bulb in Figure (27b), let *l* be the uncontracted half length of the lecture bench (6 feet for our discussion). To the Martian, that half of the lecture bench has contracted to a length  $l\sqrt{1 - v^2/c^2}$ .

In the race, the red bulb traveling at a speed v, starts out a distance  $l\sqrt{1 - v^2/c^2}$  ahead of the trigger pulse, which is traveling at a speed c. Let us assume that the race lasts a time t and that the trigger pulse catches the red bulb a distance x from where the trigger pulse started. Then we have

$$\mathbf{x} = \mathbf{ct} \tag{30}$$

In the same time t, the green bulb only travels a distance x -  $l\sqrt{1 - v^2/c^2}$ , but this must equal vt;

$$vt = x - l \sqrt{1 - v^2/c^2}$$
 (31)

Using Equation (30) in (31) gives

$$vt = ct - l\sqrt{1 - v^2/c^2}$$
  
t(c - v) =  $l\sqrt{1 - v^2/c^2}$ 

Solving for t gives

$$t = \frac{l\sqrt{1 - v^2/c^2}}{c - v} = \frac{l\sqrt{(1 + v/c)(1 - v/c)}}{c(1 - v/c)}$$
$$= \frac{l\sqrt{(1 + v/c)}}{c\sqrt{(1 - v/c)}}$$

If v is very close to c, then  $(1 + v/c) \approx 2$  and we get

$$t \approx \frac{l\sqrt{2}}{c} \left( \frac{1}{\sqrt{1 - v/c}} \right)$$

If we plug in the numbers  $t = 3 \times 10^7$  seconds (one earth year), l = 6 feet,  $c = 10^9$  feet/second, we get

$$\sqrt{1 - v/c} = \frac{l\sqrt{2}}{ct} = \frac{6 \text{ ft } \sqrt{2}}{10^9 \text{ ft/sec} \times 3 \times 10^7 \text{ sec}}$$
  
= 2.8 × 10<sup>-16</sup>

Squaring this gives

$$\left(1 - v/c\right) = 8 \times 10^{-32}$$

Thus if

$$v = (1 - 8 \times 10^{-32})c$$

=.99999999999999999999999999999992c

then *the race will last a whole year*. On the other side, the trigger signal runs into the green flash bulb in far *less than a nanosecond* because the lecture bench is highly Lorentz contracted.

# Chapter 2 Vectors

In the first chapter on Einstein's special theory of relativity, we saw how much we could learn from the simple concept of uniform motion. Everything in the special theory can be derived from (1) the idea that you cannot detect your own uniform motion, and (2) the existence of a real clock that runs slow by a factor  $\sqrt{1-v^2/c^2}$ .

We are now about to study more complicated kinds of motion where either the speed, the direction of motion, or both, are changing. Our work with non-uniform motion will be based to a large extent on a concept discovered by Galileo about 300 years before Einstein developed the special theory of relativity. It is interesting that after studying complex forms of motion for over 300 years, we still had so much to learn about simple uniform motion. But the history of science is like that. Major discoveries often occur when we see the simple underlying features after a long struggle with complex situations. If our goal is to present scientific ideas in the orderly progression from the simple to the complex, we must expect that the historical order of their discovery will not necessarily follow the same route.

Galileo was studying the motion of projectiles, trying to predict where cannon balls would land. He devised a set of experiments involving cannon balls rolling along slightly inclined planes. These experiments effectively slowed down the action, allowing Galileo to see the way the speed of a falling object changed as the object fell. To explain his results Galileo invented the concept of **acceleration** and pointed out that the simple feature of projectile motion is that projectiles move with constant or uniform acceleration. We can think of this as one step up in complexity from the uniform motion discussed in the previous chapter.

To study motion today, we have many tools that were not available to Galileo. In the laboratory we can slow down the action, or stop it, using strobe photographs or television cameras. To describe and analyze motion we have a number of mathematical tools, particularly the concept of vectors and the subject of calculus. And to predict motion, to predict not only where cannon balls land but also the trajectory of a spacecraft on a mission to photograph the solar system, we now have digital computers.

As we enter the study of more complex forms of motion, you will notice a shift in the way ideas are presented. Throughout the text, our goal is to construct a modern view of nature starting as much as possible from the basic underlying ideas. In our study of special relativity, the underlying idea, the principle of relativity, is more accurately expressed in terms of your experience flying in a jet than it is by any formal set of equations. As a result we were able to extract the content of the theory in a series of discussions that drew upon your experience.

In most other topics in physics, common experience is either not very helpful or downright misleading. If you have driven a car, you know where the accelerator pedal is located and have some idea about what acceleration is. But unless you have already learned it in a physics course, your view of acceleration will bear little relationship to the concept of acceleration developed by Galileo and now used by physicists. It is perhaps unfortunate that we use the word acceleration in physics, for we often have to spend more time dismantling the students' previous notions of acceleration than we do building the concept as used in physics. And sometimes we fail.

The physical ideas that we will study are often simply expressed in terms of mathematical concepts like a vector, a derivative, or an integral. This does not mean that we will drop physical intuition and rely on mathematics. Instead we will use them both to our best advantage. In some examples, the physical situation is obvious, and can be used to provide insight into the related mathematics. The best way, for example, to obtain a solid grip on calculus is to see it applied to physics problems. On the other hand, the concept of a vector, whose mathematical properties are easily developed, is an extremely powerful tool for explaining many phenomena in physics.

## VECTORS

In this chapter we will study the vector as a mathematical object. The idea is to have the concept of vectors in our bag of mathematical tools ready for use in our study of more complex motion, ready to be applied to the ideas of velocity, acceleration and later, force and momentum.

In a sense, we will develop a *new math* for vectors. We will begin with a definition of *displacement vectors*, and will then explain how two vectors are added. From this, we will develop a set of rules for the arithmetic of vectors. In some ways, the rules are the same as those for numbers, but in other ways they are different. We will see that most of the rules of arithmetic apply to vectors and that learning the vector convention is relatively simple.

## **Displacement Vectors**

A displacement vector is a mathematical way of expressing the separation or displacement between two objects. To see what is involved in describing the separation between objects, consider a map such as the one in Figure (1), which shows the position of the two cities, New York and Boston. If we are driving on well-marked roads, it is sufficient, when planning a trip, to know that these two cities are separated by a distance of 190 miles. However, the pilot of a small plane flying from New York to Boston in a fog must know in what direction to fly; he must also know that Boston is located at an angle of 54 degrees east of north from New York.



## Figure 1

Displacement vectors. Boston and Corning, N. Y., have equal displacements from New York and Pittsburgh, respectively. These displacements are located at different parts of the map, but they are the same displacement. The statement that Boston is located a distance of 190 miles and at an angle of 54 degrees east of north from New York provides sufficient information to allow a pilot leaving New York to reach Boston in the thickest fog. The separation or displacement between the two cities is completely described by giving both the distance and the direction.

Looking again at Figure (1), we see that Corning, N.Y., is located 190 miles, at an angle of 54 degrees east of north, from Pittsburgh. The very same instructions, travel 190 miles at an angle of 54 degrees, will take a pilot from either Pittsburgh to Corning or New York to Boston. If we say that these instructions define what we mean by the word *displacement*, then we see that Corning has the same displacement from Pittsburgh as Boston does from New York. (For our discussion we will ignore the effects of the curvature of the earth.) The displacement itself is completely described when we give both the distance and direction, and does not depend upon the point of origin.

The displacement we have been discussing can be represented graphically by an arrow pointing in the direction of the displacement (54 degrees east of north), and whose length *represents* the distance (190 mi). An arrow that represents a displacement is called a displacement vector, or simply a *vector*. One thing you should note is that a vector that defines a distance and a direction does not depend on its point of origin. In Figure (1) we have drawn two arrows; but they both represent the same displacement, and thus are *the same vector*.



#### Figure 2

Addition of vectors. The vector sum of the displacement from New York to Boston plus the displacement from Boston to Buffalo is the displacement from New York to Buffalo.

### **Arithmetic of Vectors**

Suppose that a pilot flies from New York to Boston and then to Buffalo. To his original displacement from New York to Boston he adds a displacement from Boston to Buffalo. What is the sum of these two displacements? After these displacements he will be 300 miles from New York at an angle 57 degrees west of north, as shown in Figure (2). This is the *net displacement* from New York, which is what we mean by the sum of the first two displacements.

If the pilot flies to five different cities, he is adding together five displacements, which we can represent by the vectors  $\vec{a}$ ,  $\vec{b}$ ,  $\vec{c}$ ,  $\vec{d}$ , and  $\vec{e}$  shown in Figure (3). (An arrow placed over a symbol is used to indicate that the symbol represents a vector.) Since the pilot's net displacement from his point of origin, represented by the bold vector, is simply the sum of his previous five displacements, we will say that the bold vector is the sum of the other five vectors. We will write this sum as ( $\vec{a} + \vec{b} + \vec{c} + \vec{d} + \vec{e}$ ), but remember that the addition of vectors is defined graphically as illustrated in Figure (3).

If the numbers 405 and 190 are added, the answer is 595. But, as seen in Figure (2), if you add the vector representing the 405-mile displacement from Boston to Buffalo to the vector representing the 190-mile displacement from New York to Boston, the result is a vector representing a 300-mile displacement. Clearly, there is a difference between adding numbers and vectors. The plus sign between two numbers has a different meaning from that of the plus sign between two vectors.





Although vectors differ from numbers, some similarities between the two can be noted, particularly with regard to the rules of arithmetic. First, we will review the rules of arithmetic for numbers, and then see which of these rules also apply to vectors.

## **Rules for Number Arithmetic**

1. *Commutative law*. In adding two numbers, *a* and *b*, the order of addition makes no difference.

a + b = b + a

2. *Associative law*. In adding three or more numbers, *a*, *b*, and *c*, we have

(a + b) + c = a + (b + c)

That is, if we first add *a* to *b*, and then add *c*, we get the same result as if we had added *a* to the sum (b + c).

## 3. The negative of a number is defined by

a + (-a) = 0

where (-a) is the negative of *a*.

4. *Subtraction* is defined as the addition of the negative number.

$$a - b = a + (-b)$$

These rules are so obvious when applied to numbers that it is hard to realize that they are rules. Let us apply the foregoing rules to vectors, using the method of addition of displacements.

## **Rules for Vector Arithmetic**

1. The commutative law implies that

$$\vec{a} + \vec{b} = \vec{b} + \vec{a}$$

Figure (4) verifies this rule graphically. The reader should be able to see that  $\vec{a} + \vec{b}$  and  $\vec{b} + \vec{a}$  are the same vectors.



Figure 4

2. The associative law applied to vectors would imply

$$(\vec{a} + \vec{b}) + \vec{c} = \vec{a} + (\vec{b} + \vec{c})$$

From Figure (5) you should convince yourself that this law works.



Figure 5

#### 3. The negative of a vector is defined by

$$\vec{a} + \vec{-a} = 0$$

The only way to get a *zero displacement* is to return to the point of origin. Thus, the negative of a vector is a vector of the same length but pointing in the opposite direction (Figure 6).

Figure 6

**4.** The subtraction of vectors is now easy. If we want  $\vec{a} \cdot \vec{b}$ , we just find  $\vec{a} + \vec{b}$ . That is

 $\vec{a} - \vec{b} = \vec{a} + \vec{b}$ 

To subtract, we just add the negative vector as shown in Figure (7).



Figure 7

## Multiplication of a Vector by a Number

Suppose we multiply a vector  $\vec{a}$  by the number 5. What do we mean by the result  $5\vec{a}$ ? Let us again try to follow the rules of arithmetic to answer this question. In arithmetic we were taught that

5a=a+a+a+a+a

Let us try the same rule for vectors.

$$5\vec{a} = \vec{a} + \vec{a} + \vec{a} + \vec{a} + \vec{a}$$

With this definition we see that  $5\vec{a}$  is a vector in the same direction as  $\vec{a}$  but five times as long (see Figure 8).



#### Figure 8

We may also multiply a vector by a negative number (see Figure 9); the minus sign just turns the vector around. For example,

$$-3\vec{a} = 3(-\vec{a}) = -\vec{a} + -\vec{a} + -\vec{a}$$



Figure 9

When we multiply a vector by a positive number, we merely change the length of the vector; multiplication by a negative number changes the length *and* reverses the direction.

## **Magnitude of a Vector**

Often we will want to discuss only the length or magnitude of a vector, regardless of the direction in which it is pointing. For example, if we represent the displacement of Boston from New York by the vector  $\vec{s}$ , then the magnitude of  $\vec{s}$  (the length of this displacement) is 190 miles. We use a vertical bar on each side of the vector to represent the magnitude ; thus, we write  $|\vec{s}| = 190$  mi (see Figure 10).



## **Vector Equations**

Just as we can solve algebraic equations involving numbers, we can do the same for vectors. Suppose, for example, we would like to find the vector  $\vec{x}$  in the vector equation

 $2\vec{a} + 3\vec{b} + 2\vec{x} = \vec{c}$ 

Solving this equation the same way we would any other, we get

 $\vec{x} = 1/2\vec{c} - \vec{a} - 3/2\vec{b}$ 

Graphically, we find  $(1/2)\vec{c}$ ,  $-\vec{a}$ , and  $(-3/2)\vec{b}$ ; we then vectorially add these quantities together to get the vector  $\vec{x}$ .

## **Graphical Work**

In the early sections of this text, we shall do a fair amount of graphical work with vectors. As we can see from the previous examples, the main problem in graphical work is to move a vector accurately from one part of the page to another. This is easily done with a plastic triangle and ruler as described in the following example.

#### Example 1

b•

The vector  $\vec{s}$  starts from point *a* and we would like to redraw it starting from point *b*, as shown in Figure (11).



**Solution:** We want to draw a line through *b* that is parallel to  $\vec{s}$ . This can be done with a straightedge and triangle as shown in Figure (12).



Place the straight edge and triangle so that one side of the triangle lies along the straight edge and the other along the vector  $\vec{s}$ . Then slide the triangle along the straight edge until the side of the triangle that was originally along  $\vec{s}$  now passes through *b*. Draw this line through b. If nothing has slipped, the line will be parallel to  $\vec{s}$  as shown in Figure (13).



We now have the direction of  $\vec{s}$  starting from b. Thus, we have only to put in the length. This is most easily done by marking the length of  $\vec{s}$  on the edge of a piece of paper and reproducing this length, starting from b as shown in Figure (14).



By being careful, using a sharp pencil, and practicing, you should have no difficulty in performing accurate and rapid graphical work. The practice can be gained by doing Problems 1 through 5. (Note that it is essential to distinguish a vector from a number. Therefore, when you are solving problems or working on a laboratory experiment, *it is recommended that you always place an arrow over the symbol representing a vector.*)

#### Exercise 1 Commutative law

The vectors  $\vec{a}$ ,  $\vec{b}$ , and  $\vec{c}$  of Figure 15 are shown enlarged on the tear out page 2-19. Using that page for your work, find



#### Exercise 2 Associative law

Use the tear out page 2-20 for the vectors of Figure (16), find



#### Exercise 3 Subtraction

Use the tear out page 2-21 for the three vectors  $\vec{a}$ ,  $\vec{b}$ , and  $\vec{c}$  shown in Figure (17), find the following vectors graphically, labeling your results.



## Exercise 4 Equations

Suppose that a physical law is given by the vector equation

$$\vec{P}_i = \vec{P}_f$$

Suppose that  $\vec{P}_{f}$  is the sum of two vectors; that is,

$$\vec{P}_{f} = \vec{P}_{f1} + \vec{P}_{f}$$

Given the two vectors  $P_i$  and  $\overrightarrow{P_{f1}}$  (Figure 18), find  $\overrightarrow{P_{f2}}$ . (These vectors are found on the tear out page 2-22.)



### Exercise 5

Assume that the vectors  $\vec{P_{f1}}$ ,  $\vec{P_{f1}}$ , and  $\vec{P_{f2}}$  are related by the *vector* law:

$$\vec{P}_{f} = \vec{P}_{f1} + \vec{P}_{f2}$$

In addition, the magnitudes of the vectors are related by

 $\left| \mathsf{P}_{\mathsf{f}} \right|^2 = \left| \mathsf{P}_{\mathsf{f}\,1} \right|^2 + \left| \mathsf{P}_{\mathsf{f}\,2} \right|^2$ 

If you are given  $\vec{P_f}$  and only the *direction* of  $\vec{P_{f1}}$  (Figure 19), find  $\vec{P_{f1}}$  and  $\vec{P_{f2}}$  graphically. (These vectors are found on the tear out page 2-22.)





**Figure 20** *Two ways to reach Boston from New York.* 

## COMPONENTS

Another way to work with vectors, one that is especially convenient for solving numerical problems, is through the use of a coordinate system and components. To illustrate this method, suppose we were giving instructions to a pilot on how to fly from New York to Boston. One way, which we have mentioned, would be to tell the pilot both the direction and the distance she must fly, as "fly at an angle of 54 degrees east of north for a distance of 190 miles." But we could also tell her "fly 132 miles due east and then fly 112 miles due north." This second routing, which describes the displacement in terms of its easterly and northerly components, as illustrated in Figure (20), is less direct, but will also lead the pilot to Boston.

We can use the same alternate technique to describe a vector drawn on a piece of paper. In Figure (20), we drew two lines to indicate easterly and northerly directions. We have drawn the same lines in Figure (21), but now we will say that these lines represent the *x* and *y* directions. The lines themselves are called the *x* and *y* axes, respectively, and form what is called a coordinate system.

Just as the displacement from New York to Boston had both an easterly and northerly component, the vector  $\vec{a}$ in Figure (21) has both an *x* and *a y* component. In fact, the vector  $\vec{a}$  is just the sum of its component vectors  $\vec{a}_x$ and  $\vec{a}_y$ :

$$\vec{a} = \vec{a}_x + \vec{a}_y \tag{1}$$



**Figure 21** Component vectors. The sum of the component vectors  $\vec{a}_x$  and  $\vec{a}_y$  is equal to the vector  $\vec{a}$ .

Trigonometry can be used to find the length or magnitude of the component vectors; we get

$$\mathbf{a}_{\mathbf{x}} \equiv \left| \overrightarrow{\mathbf{a}}_{\mathbf{x}} \right| = \left| \overrightarrow{\mathbf{a}} \right| \cos \theta \tag{2}$$

$$a_{y} \equiv \left| \overrightarrow{a_{y}} \right| = \left| \overrightarrow{a} \right| \sin \theta \tag{3}$$

Often we will represent the magnitude of a component vector by not using the arrow, as was done in the foregoing equations. (The equal sign with three bars,  $a_x \equiv |\vec{a_x}|$ , simply means that  $a_x$  *is defined to be the same symbol as*  $|\vec{a_x}|$ .) It is common terminology to call the magnitude of a component vector simply the *component;* for example,  $a_x \equiv |\vec{a_x}|$  is called *the x component of the vector d*.

#### Addition of Vectors by Adding Components

An important use of components is as a means for handling vectors numerically rather than graphically. We will show how this works by using an example of the addition of vectors by adding components.

Consider the three vectors shown in Figure (22). Since each vector is the vector sum of its individual components vectors, we have

$$\vec{a} = \vec{a_x} + \vec{a_y}$$
  $\vec{b} = \vec{b_x} + \vec{b_y}$   $\vec{c} = \vec{c_x} + \vec{c_y}$ 

By adding all three vectors  $\vec{a}$ ,  $\vec{b}$ , and  $\vec{c}$  together, we get

$$\vec{a} + \vec{b} + \vec{c} = (\vec{a_x} + \vec{a_y}) + (\vec{b_x} + \vec{b_y}) + (\vec{c_x} + \vec{c_y})$$

The right-hand side of this equation may be rearranged to give



Equation 4 gives us a new way to add vectors, as illustrated in Figure (23). Previously we would have added the vectors  $\vec{a}$ ,  $\vec{b}$ , and  $\vec{c}$  directly, as shown in Figure (24). The new rule shows how we can first add the x components  $(\vec{a_x} + \vec{b_x} + \vec{c_x})$  as shown in Figure (23a), then separately add the y components  $(\vec{a_y} + \vec{b_y} + \vec{c_y})$  as shown in Figure (23b), and then add these vector sums vectorially, as shown in Figure (23c), to get the vector  $(\vec{a} + \vec{b} + \vec{c})$ .

(a) 
$$\frac{\vec{a}_{x} \qquad \vec{b}_{x}}{(\vec{a}_{x} + \vec{b}_{x} + \vec{c}_{x}) \qquad \vec{c}_{x}}$$

(b) 
$$(\vec{a}_y + \vec{b}_y + \vec{c}_y) \vec{c}_y$$

$$(\vec{a}_{x} + \vec{b}_{x} + \vec{c}_{x})$$
$$(\vec{a}_{y} + \vec{b}_{y} + \vec{c}_{y})$$

Figure 23

(c)



Figure 24

The advantage of using components is that we can numerically add or subtract the lengths of vectors that point in the same direction. Thus, to add 500 vectors, we would compute the lengths of all the x components and add (or subtract) these together. We would then add the lengths of the y components, and finally, we would vectorially add the resulting x and y components. Since the x and y components are at right angles, we may find the total length and final direction by using the Pythagorean theorem and trigonometry, as shown in Figure (25).



It is not necessary to always choose the x components horizontally and the y components vertically. We may choose a coordinate system (x', y') tilted at an angle, as shown in Figure (26). To use the language of the mathematician,  $a_{x'}$  is the component of (or projection of)  $\vec{a}$  in the direction x'. We see that the vector sum of all the component vectors still adds up to the vector itself.



Figure 26

#### **Exercise 6**

Imagine you are given the vectors  $\vec{a}$ ,  $\vec{b}$ , and  $\vec{c}$  and the two sets of coordinate axes  $(x_1, y_1)$  and  $(x_2, y_2)$  shown in Figure (27). Using the vectors found on the tear out page 2-23



a) Find  $(\vec{a} + \vec{b} + \vec{c})$  by direct addition of vectors.

b) Choose  $x_1$  and  $y_1$  as your coordinate axes. Find (in red) the  $x_1$  and  $y_1$  components of  $\vec{a}, \vec{b}, \vec{c}$ . Then

(i) Find  $(\vec{a}_{x1} + \vec{b}_{x1} + \vec{c}_{x1})$ (ii) Find  $(\vec{a}_{y1} + \vec{b}_{y1} + \vec{c}_{y1})$ (iii) Find  $(\vec{a}_{x1} + \vec{b}_{x1} + \vec{c}_{x1}) + (\vec{a}_{y1} + \vec{b}_{y1} + \vec{c}_{y1})$ .

How does this compare with  $(\vec{a} + \vec{b} + \vec{c})$ ?

c) Repeat part (b) for the coordinate axis  $(x_2, y_2)$ .

## **Vector Equations in Component Form**

Often we will run into a situation where we have a vector equation of the form

$$\vec{c} = \vec{a} + \vec{b}$$

but you have to solve the equation using components. This is easy to do, because to go from a vector equation to component equations, just rewrite the equation three (or two) times, once for each component. The above equation becomes

$$c_x = a_x + b_x$$
  

$$c_y = a_y + b_y$$
  

$$c_z = a_z + b_z$$

## **VECTOR MULTIPLICATION**

We have seen how the rules work for vector addition, subtraction, and the multiplication of a vector by a number. Does it make any sense to multiply two vectors together? In considering the multiplication of the two vectors, the first question to answer is: what is the result? What kind of a thing do we get if we multiply a vector pointing east by a vector pointing north? Do we get a vector pointing in some third direction? Do we get a number that does not point? Or do we get some quantity more complex than a vector? And perhaps a more important question – why would one want to multiply two vectors together?

We will see in the study of physics that there are various reasons why we will want to multiply vectors, and we can get various answers. One kind of multiplication produces a number; this is called scalar multiplication or the *dot product*. We will see examples of scalar multiplication shortly. A few chapters later we will encounter the *vector cross product* where the result of the multiplication of two vectors is itself a vector, one that points in a direction perpendicular to the two vectors being multiplied together. Finally there is a form of multiplication that leads to a quantity more complex than a vector, an object called a *tensor* or a matrix. A tensor is an object that maintains the directional nature of both vectors involved in the product. Tensors are useful in the formal mathematical description of the basic laws of physics, but are not needed and will not be used in this text.

The names *scalar*, *vector*, and *tensor* describe a hierarchy of mathematical quantities. Scalars are numbers like, 1, 3, and -7, that have a magnitude but do not point anywhere. Vectors have both a magnitude and a direction. Tensors have the basic properties of both vectors used to construct them. In fact there are higher rank tensors that have the properties of 3, 4, or more vectors. People working with Einstein's generalized gravitational theory have to work all the time with tensors. One of the remarkable discoveries of the twentieth century is that there is a close relationship between the mathematical properties of scalars, vectors, and tensors, and the physical properties of the various elementary particles. Later on we will discuss particles such as the  $\pi$  meson now used in cancer research, the photon which is the particle of light (a beam of light is a beam of photons), and the graviton, the particle hypothesized to be responsible for the gravitational force. It turns out that the physical properties of the  $\pi$  meson resemble the mathematical properties of a scalar, the properties of the photon are described by a vector (we will see this later in the text), and it requires a tensor to describe the graviton (that is why people working with gravitational theories have to work with tensors).

One of the surprises of physics and mathematics is that there are particles like the electron, proton and neutron, the basic constituents of atoms, that are not described by scalars, vectors, or tensors. To describe these particles, a new kind of a mathematical object had to be invented-an object called the spinor. The spinor describing the electron has properties half way between a scalar and a vector. No one knew about the existence of spinors until the discovery was forced by the need to explain the behavior of electrons. In this text we will not go into the mathematics of spinors, but we will encounter some of the unusual properties that spinors have when we study the behavior of electrons in atoms. In a very real sense the spinor nature of electrons is responsible for the periodic table of elements and the entire field of chemistry.

In this text we can discuss a great many physical concepts using only scalars or vectors, and the two kinds of vector products that give a scalar or vector as a result. We will first discuss the scalar or dot product which is some ways is already a familiar concept, and then the vector or cross product which plays a significant role later in the text.

#### **The Scalar or Dot Product**

In a scalar product, we start with two vectors, multiply them together, and get a number as a result. What kind of a mathematical process does that involve? The Pythagorean theorem provides part of the answer.

Suppose that we have a vector  $\vec{a}$  whose x and y components are  $a_x$  and  $a_y$  as shown in Figure (28). Then the magnitude or length  $|\vec{a}|$  of the vector is given by the Pythagorean theorem as

$$|\vec{a}|^2 = (a_x)^2 + (a_y)^2 \tag{4}$$





In some sense  $|\vec{a}|^2$  is the product of the vector  $\vec{a}$  with itself, and the answer is a number that is equal to the square of the length of the vector  $\vec{a}$ .

Now suppose that we use a different coordinate system x', y' shown in Figure (29) but have the same vector  $\vec{a}$ . In this new coordinate system the length of the vector  $\vec{a}$  is given by the formula



The components  $a'_x$  and  $a'_y$  are different from  $a_x$  and  $a_y$ , but we know that the length of  $|\vec{a}|$  has not changed, thus  $|\vec{a}|^2$  must be the same in Equations (4) and (5). We have found a quantity  $|\vec{a}|^2$  which has the same value in all coordinate systems even though the pieces  $a_x^2$  and  $a_y^2$  change from one coordinate system to another. This is the key property of what we will call the scalar product.

To formalize this concept, we will define the *scalar product* of the vector  $\vec{a}$  with itself as being the square of the length of  $|\vec{a}|$ . We will denote the scalar product by using the *dot* symbol to denote scalar multiplication:

Scalar product  
of 
$$\vec{a}$$
 with itself  $\equiv \vec{a} \cdot \vec{a} \equiv |\vec{a}|^2$  (6)

From Equations (4) and (6) we have in the (x, y) coordinate system

$$\vec{a} \cdot \vec{a} = a_x^2 + a_y^2 \tag{7}$$

In the (x', y') coordinate system we get

$$\vec{a} \cdot \vec{a} = \left(a'_{x}\right)^{2} + \left(a'_{y}\right)^{2}$$
(8)

The fact that the length of the vector  $\vec{a}$  is the same in both coordinate systems means that this scalar or "dot" product of  $\vec{a}$  with itself has the same value even though the components or pieces  $a_x^2$ ,  $a_y^2$  or  $(a'_x)^2$ ,  $(a'_y)^2$  are different. In a more formal language, we can say that the scalar product  $\vec{a} \cdot \vec{a}$  is unchanged by, or *invariant* under changes in the coordinate system. Basically we can say that there is physical meaning to the quantity  $\vec{a} \cdot \vec{a}$  (i.e. the length of the vector) that does not depend upon the coordinate system used to measure the vector.

#### **Exercise 7**

Find the dot product  $\vec{a} \cdot \vec{a}$  for a vector with components  $a_x, a_y, a_z$  in three dimensional space. How does the Pythagorean theorem enter in this case?

The example of calculating  $\vec{a} \cdot \vec{a}$  above gives us a clue to guessing a more general definition of dot or scalar products when we have to deal with the product of two different vectors  $\vec{a}$  and  $\vec{b}$ . As a guess let us try as a definition

$$\vec{a} \cdot \vec{b} \equiv a_x b_x + a_y b_y \tag{9}$$

or in three dimensions

$$\vec{a} \cdot \vec{b} \equiv a_x b_x + a_y b_y + a_z b_z$$
(10)

This definition of a dot product does not represent the length of either  $\vec{a}$  or  $\vec{b}$  but perhaps  $\vec{a} \cdot \vec{b}$  has the special property that its value is independent of the choice of coordinate system, just as  $\vec{a} \cdot \vec{a}$  had the same value in any coordinate system. To find out we need to calculate the quantity  $(a'_x b'_x + a'_y b'_y + a'_z b'_z)$  in another coordinate system and see if we get the same answer. We will do a simple case to show that this is true, and leave the more general case to the reader.



Suppose we have two vectors  $\vec{a}$  and  $\vec{b}$  separated by an angle q as shown in Figure (30). Let the lengths  $|\vec{a}|$  and  $|\vec{b}|$  be denoted by a and b respectively. Choosing a coordinate system (x, y) where the x axis lines up with  $\vec{a}$ , we have

$$a_{x} = a, a_{y} = 0$$
  
$$b_{x} = b \cos \theta, b_{y} = b \sin \theta$$

and the dot product, Equation 9, gives

$$(\vec{a} \cdot \vec{b}) = a_x b_x + a_y b_y$$
  
=  $ab \cos \theta + 0$ 



#### Figure 31

Next choose a coordinate system (x',y') rotated from (x,y) by an angle  $(90 - \theta)$  as shown in Figure (31). Here b lies along the y'axis and the dot product is given by

$$\vec{a} \cdot \vec{b} = a'_x b'_x + a'_y b'_y$$
  
=  $ab \cos \theta + 0$ 

Again we get the result

$$\vec{a} \cdot \vec{b} = ab \cos \theta \tag{11}$$

Equation (11) holds no matter what coordinate system we use, as you can see by working the following exercise.

#### Exercise 8

Choose a coordinate system (x'', y'') where the x axis is an angle  $\phi$  below the horizontal as shown in Figure (32). First calculate the components  $a''_{x}, a''_{y}, b''_{x}, b''_{v}$  and then show that you still get

$$\vec{a} \cdot \vec{b} \equiv a_x'' b_x'' + a_y' b_y'' = ab\cos\theta$$

Figure 32



To do this problem, you need the following relationships.

$$\sin(\theta + \phi) = \sin(\theta)\cos(\phi) + \cos(\theta)\sin(\phi)$$
$$\cos(\theta + \phi) = \cos(\theta)\cos(\phi) - \sin(\theta)\sin(\phi) )$$
$$\sin^{2}(\phi) + \cos^{2}(\phi) = 1 \qquad \text{for any angle } \phi$$

(This problem is much messier than the example we did.)

#### Interpretation of the Dot Product

When  $\vec{a}$  and  $\vec{b}$  are the same vector, then we had  $\vec{a} \cdot \vec{a} = |\vec{a}|^2$  which is just the square of the length of the vector. If  $\vec{a}$  and  $\vec{b}$  are different vectors but parallel to each other, then  $\theta = 0^\circ$ ,  $\cos \theta = 1$ , and we get

$$\vec{a} \cdot \vec{b} = ab$$

In other words the dot product of parallel vectors is just the product of the lengths of the vectors.

Another extreme is when the vectors are perpendicular to each other. In this case  $\theta = 90^{\circ}$ ,  $\cos \theta = 0$ and  $\vec{a} \cdot \vec{b} = 0$ . The dot product of perpendicular vectors is zero. In a sense



the dot product of two vectors measures the parallelism of the vectors. If the two vectors are parallel, the dot product is equal to the full product ab. If they are perpendicular, we get nothing. If they are at some intermediate angle, we get a number between ab and zero.

Increasing  $\theta$ more, we see that if the vectors  $\vec{b}$ are separated by an angle between 90° and 180° as in Figure (33), then the cos  $\theta$  and the dot product are negative. A negative dot product indicates an anti-parallelism. The extreme case is  $\theta = 180^\circ$ where  $\vec{a} \cdot \vec{b} = -ab$ .



#### **Physical Use of the Dot Product**

We have seen that the dot product  $\vec{a} \cdot \vec{b}$  is given by the simple formula  $\vec{a} \cdot \vec{b} = a b \cos \theta$  and it has the special property that

$$\vec{a} \cdot \vec{b} \equiv a_x b_x + a_y b_y + a_z b_z$$

has the same value in any coordinate system even though the components  $a_x, b_x$  etc., are different in different coordinate systems. The fact that  $\vec{a} \cdot \vec{b}$  is the same number in different coordinate systems means that it is truly a number with no dependence on direction. That is what we mean by a *scalar quantity*. This is a special property because  $\vec{a} \cdot \vec{b}$  is made up of the vectors  $\vec{a}$  and  $\vec{b}$  that *do* depend upon direction and whose values *do* change when we go to different coordinate systems.

In physics there are quantities like displacements  $\vec{x}$ , velocities  $\vec{v}$ , forces  $\vec{F}$  that all behave like vectors. All point somewhere and have components that depend upon our choice of direction. Yet we will deal with other quantities like energy which does not point anywhere. Energy has a magnitude but no direction. Yet our formulas for energy involve the vectors  $\vec{x}$ ,  $\vec{v}$ , and  $\vec{F}$ . How can we construct numbers or scalars from vectors? The answer is - take dot or scalar products of the vectors. This is the mathematical reason why most of our formulas for energy will involve dot products.

## **Vector Cross Product**

The other kind of vector product we will use in this course is the vector cross product where we multiply two vectors  $\vec{a}$  and  $\vec{b}$  together to get a third vector  $\vec{c}$ . The notation is

$$\vec{a} \times \vec{b} = \vec{c} \tag{12}$$

where the name *cross product* comes from the cross we place between the vectors we are multiplying together. When you first encounter the cross product, it does not seem particularly intuitive. But we use it so much in later chapters that you will get quite used to it. Perhaps the best procedure is to skim over this material now, and refer back to it later when we start using it in various physics applications.

To define the cross product  $\vec{a} \times \vec{b} = \vec{c}$ , we have to define not only the magnitude but also the direction of the resulting vector  $\vec{c}$ . Starting with two vectors  $\vec{a}$  and  $\vec{b}$  pointing



in different directions as in Figure 34, what unique direction is there for  $\vec{c}$  to point? Should  $\vec{c}$  point half way between  $\vec{a}$  and  $\vec{b}$ , or should it be closer to  $\vec{a}$  because  $\vec{a}$  is

longer than b? No, there is nothing particularly unique or obvious about any of the directions in the plane defined by  $\vec{a}$ and  $\vec{b}$ . The only truly unique direction is perpendicular to this plane. We will say that  $\vec{c}$ points in this unique direction as shown in Figure 35.



The direction perpendicular to the plane of  $\vec{a}$  and  $\vec{b}$  is not quite unique. The vector  $\vec{c}$  could point either up or down as indicated by the solid or dotted vector in Figure 35. To select between these two choices, we use what is called the *right hand rule* which can be stated as follows: Point the fingers of your right hand in the direction of the first of the two vectors in the cross product  $\vec{a} \times \vec{b}$  (in this case the vector  $\vec{a}$ ). Then curl your fingers until they point in the direction of the second vector (in this case  $\vec{b}$ ), as shown in Figure 36. If you orient your right hand so that this curling is physically possible, then your thumb will point in the direction of the cross product vector  $\vec{c}$ .

#### **Exercise 9**

What direction would the vector  $\vec{c}$  point if you used your left hand rather than your right hand in the above rule?

We said that the vector cross product was not a particularly intuitive concept when you first encounter it. In the above exercise, you see that if by accident you use your left hand rather than your right hand,  $\vec{c} = \vec{a} \times \vec{b}$ will point the other way. One can reasonably wonder how a cross product could appear in any law of physics, for why would nature prefer right hand rules over left handed rules. It seems unbelievable that any basic concept should involve anything as arbitrary as the right hand rule.

There are two answers to this problem. One is that in most cases, nature has no preference for right handedness over left handedness. In these cases it turns out that any law of physics that involves right hand rules turns out to involve an even number of them so that any physical prediction does not depend upon whether you used a right hand rule or a left hand rule, as long as you use the same rule throughout. Since there are more right handed people than left handed people, the right hand rule has been chosen as the standard convention.



**Figure 36** *Right hand rule for the vector cross product.* 

### **Exercise 10**

There is left and right handedness in the direction of the threads on a screw or bolt. In Figure (37a) we show a screw with a right handed thread. By this, we mean that if we turn the screw in the direction that we can curl the fingers of our right hand, the screw will move through wood in the direction that the thumb of our right hand points.



**Figure 37a** *Right handed thread.* 

In Figure (37b), we have a left hand thread. If we turn the screw in the direction we can curl the fingers of our left hand, the screw will move through the direction pointed by our left thumb.



**Figure 37b** Left handed thread

For this exercise find some screws and bolts, and determine whether the threads are right handed or left handed. Manufacturers use one kind of thread predominately over the other. Which is the predominant thread? Can you locate examples of the other kind of thread? (The best place to look for the other kind of thread is in the mechanism of some water faucets. Can you find a water faucet where one side uses a right hand thread and the other a left hand thread? If you find one, determine which is the right and which the left hand thread.) Until 1956 it was believed that the basic laws of physics did not distinguish between left and right handedness. The fact that there are more right handed than left handed people, or that the DNA used by living organisms had a right handed spiral structure (like a right handed thread) was simply an historical accident. But then in 1956 it was discovered that the elementary particle called the *neutrino* was fundamentally left handed. Neutrinos spin like a top. If a neutrino is passing by you and you point the thumb of your left hand in the direction the neutrino is moving, the fingers of your left hand curl in the direction that the neutrino is spinning. Or we may say that the neutrino turns in the direction of a left handed thread, as shown in Figure 38.



#### Figure 38

The neutrino is inherently a left handed object. When one passes by you, it spins in the direction that the threads on a left handed screw turn.

Another particle, called the anti-neutrino, is right handed. If you point the thumb of your right hand in the direction of motion of an anti-neutrino, the fingers of your right hand can curl in the direction that the antineutrino rotates. T.D. Lee and N.C. Yang received the 1957 Nobel prize in physics for their discovery that some basic phenomena of physics can be used to distinguish between left and right handedness.

The idea of right or left handedness in the laws of physics will appear in several of our later discussions of the basic laws of physics. The point for now is that having a quantity like the vector cross product that uses the right hand convention may be a useful tool to distinguish between left and right handedness.

#### Exercise 11

Go back to Figure 34 where we show the vectors  $\vec{a}$  and  $\vec{b}$ , and draw the vector  $\vec{c}' = \vec{b} \times \vec{a}$ . Use the right hand rule as we stated it to determine the direction of  $\vec{c}'$ . From your result, decide what happens when you reverse the order in which you write the vectors in a cross product. Which of the arithmetic rules does this violate?

## **Magnitude of the Cross Product**

Now that we have the right hand rule to determine the direction of  $\vec{c} = \vec{a} \times \vec{b}$ , we now need to specify the magnitude of  $\vec{c}$ .



#### Figure 39

A clue as to a consistent definition of the magnitude of  $\vec{c}$  is the fact that when  $\vec{a}$  and  $\vec{b}$  are parallel, they do not define a plane. In this special case there is an entire plane perpendicular to both  $\vec{a}$  and  $\vec{b}$ , as shown in Figure 39. Thus there is an infinite number of directions that  $\vec{c}$  could point and still be perpendicular to both  $\vec{a}$  and  $\vec{b}$ . We can avoid this mathematical ambiguity only if  $\vec{c}$  has zero magnitude when  $\vec{a}$  and  $\vec{b}$  are parallel. We do not care where  $\vec{c}$  points if it has no length.



Figure 40

## The simplest formula for the magnitude $|\vec{c}| = |\vec{a} \times \vec{b}|$ , that is related to the product of $\vec{a}$ and $\vec{b}$ , yet has zero length when $\vec{a}$ and $\vec{b}$ are parallel is

$$|\vec{c}| = |\vec{a} \times \vec{b}| = ab \sin\theta$$
 (13)

where  $a \equiv |\vec{a}|$  and  $b \equiv |\vec{b}|$  are the lengths of  $\vec{a}$  and  $\vec{b}$  respectively, and  $\theta$  is the angle between them. Equation 13 is the definition we will use for the magnitude of the vector cross product.

In Equation 13, we see that not only is the cross product zero when the vectors are parallel, but is a maximum when the vectors are perpendicular. In the sense that the dot product  $\vec{a} \cdot \vec{b}$  was a measure of the parallelism of the vectors  $\vec{a}$  and  $\vec{b}$ , the cross product is a measure of their perpendicularity. If  $\vec{a}$  and  $\vec{b}$  are perpendicular, then the length of c is just the product ab. As the vectors become parallel the length of c reduces to zero.

## Component Formula for the Cross Product

Sometimes one needs the formula for the components of  $\vec{c} = \vec{a} \times \vec{b}$  expressed in terms of the components of  $\vec{a}$  and  $\vec{b}$ . The result is a mess, and is remembered only by those who frequently use cross products. The answer is

$$c_{x} = a_{y}b_{z} - a_{z}b_{y}$$

$$c_{y} = a_{z}b_{x} - a_{x}b_{z}$$

$$c_{z} = a_{x}b_{y} - a_{y}b_{x}$$
(14)

These formulas are not so bad if you are doing a computer calculation and you are letting the computer evaluate the individual components.

#### **Exercise 12**

Assume that  $\vec{a}$  points in the x direction and  $\vec{b}$  is in the xy plane as shown in Figure 41. By the right hand rule,  $\vec{c}$  will point along the z axis as shown. Use Equation 14 to calculate the magnitude of  $c_z$  and compare your result with Equation 13.





## RIGHT HANDED COORDINATE SYSTEM

Notice in Figure 41, we have drawn an (x, y, z) coordinate system where z rises up from the xy plane. We could have drawn z down and still have three perpendicular directions. Why did we select the upward direction for z?

The answer is that the coordinate system shown in Figure 41 is a **right hand coordinate system**, defined as follows. Point the fingers of your right hand in the direction of the first coordinate axis(x). Then curl your fingers toward the second coordinate axis(y). If you have oriented your right hand so that you can curl your fingers this way, then your thumb points in the direction of the third coordinate axis(z).

The importance of using a right handed coordinate system is that Equation 14 for the cross product expressed as components works only for a right handed coordinate system. If by accident you used a left handed coordinate system, the signs in the equation would be reversed.

## Exercise 13

Decide which of the (x, y, z) coordinate systems are right handed and which are left handed.



**Figure 15** *Vectors for Exercise 1, page 7. Find* 

- (a)  $\vec{a} + \vec{b} + \vec{c}$  (in black);
- (b)  $\vec{b} + \vec{c} + \vec{a}$ (in red);
- (c)  $\vec{c} + \vec{a} + \vec{b}$ (in blue).



## Figure 16

## Vectors for Exercise 2, page 7. Find

- (a)  $\vec{a} + \vec{b}$  (in black); (b)  $(\vec{a} + \vec{b}) + \vec{c}$  (in red);
- (c)  $(\vec{b} + \vec{c})$  (in black);
- (d)  $\vec{a} + (\vec{b} + \vec{c})$  (in blue).



**Figure 17** *Vectors for Exercise 3, page 7. Find* 

- (a)  $\vec{a} + \vec{b}$
- (b)  $\vec{a} \vec{b}$
- (c)  $\vec{b} \vec{a}$
- (d)  $(\vec{a} \vec{b}) + (\vec{b} \vec{a})$
- (e)  $\vec{b} + \vec{c} \vec{a}$



## Figure 18

Vectors for Exercise 4, page 7.

$$P_i = P_f$$

Suppose that  $\overrightarrow{P_f}$  is the sum of two vectors; that is,

$$\overrightarrow{P_f} = \overrightarrow{P_{f1}} + \overrightarrow{P_{f2}}$$

Given the two vectors  $\overrightarrow{P_{i}}$  and  $\overrightarrow{P_{f1}}$  (Figure 18), find  $\overrightarrow{P_{f2}}$ 



Figure 19 Vectors for Exercise 5, page 8.  $\overrightarrow{P_{f}} = \overrightarrow{P_{f1}} + \overrightarrow{P_{f2}}$ 

In addition, the magnitudes of the vectors are related by

$$\left|\overrightarrow{\mathbf{P}_{f}}\right|^{2} = \left|\overrightarrow{\mathbf{P}_{f1}}\right|^{2} + \left|\overrightarrow{\mathbf{P}_{f2}}\right|^{2}$$

If you are given  $\vec{P}_f$  and only the direction of  $\overrightarrow{P_{f1}}$  , find  $\overrightarrow{P_{f1}}$  and  $\overrightarrow{P_{f2}}$  graphically.



## Figure 27

## Vectors for Exercise 6, page 10.

- (a) Find  $\vec{a} + \vec{b} + \vec{c}$  by direct addition of vectors.
- (b) Choose  $x_1$  and  $y_1$  as your coordinate axes. Find (in red) the  $x_1$  and  $y_1$  components of  $\vec{a},\,\vec{b},\,\vec{c}.$  Then
  - (i) Find  $\vec{a}_{x1} + \vec{b}_{x1} + \vec{c}_{x1}$
  - (ii) Find  $\vec{a}_{y1} + \vec{b}_{y1} + \vec{c}_{y1}$
  - (iii) Find  $(\vec{a}_{x1} + \vec{b}_{x1} + \vec{c}_{x1}) + (\vec{a}_{y1} + \vec{b}_{y1} + \vec{c}_{y1}).$

How does this compare with  $(\vec{a} + \vec{b} + \vec{c})$ ?

(c) Repeat part B for the coordinate axis  $(x_2, y_2)$ . (you can use the back side of this page.)




# Figure 27

# Vectors for Exercise 6, page 10, repeated.

- (a) Find  $\vec{a} + \vec{b} + \vec{c}$  by direct addition of vectors.
- (b) Choose  $x_1$  and  $y_1$  as your coordinate axes. Find (in red) the  $x_1$  and  $y_1$  components of  $\vec{a},\,\vec{b},\,\vec{c}.$  Then
  - (i) Find  $\vec{a}_{x1} + \vec{b}_{x1} + \vec{c}_{x1}$
  - (ii) Find  $\vec{a}_{y1} + \vec{b}_{y1} + \vec{c}_{y1}$

(iii) Find 
$$(\vec{a}_{x1} + \vec{b}_{x1} + \vec{c}_{x1}) + (\vec{a}_{y1} + \vec{b}_{y1} + \vec{c}_{y1})$$
.

How does this compare with  $(\vec{a} + \vec{b} + \vec{c})$ ?

(c) Repeat part B for the coordinate axis  $(x_2, y_2)$ .







**Figure 1** Marcel Duchamp, *Nude Decending a Staircase* Philadelphia Museum of Art: Louise and Walter Arensberg Collection

# Chapter 3 Description of Motion

On the facing page is a reproduction (Figure 1) of Marcel Duchamp's painting, Nude Descending a Staircase, which was first displayed in New York at The International Exhibition of Modern Art, generally known as the Armory Show, in 1913. The objective of the painting, to convey a sense of motion, is achieved by repeating the stylized human form five times as it descends the steps. At the risk of obscuring the artistic qualities of the painting, we may imagine this work as a series of five flash photographs taken in sequence as the model walked downstairs. In the next few chapters, a similar technique will be used to describe motion. We now have devices available, such as the stroboscope (called the strobe), that produce short bursts of light at regular intervals; with the strobe, we can photograph the successive positions of an object, such as a ball moving on the end of a string (see Figure 2). Although we do not have the artist's freedom of expression to convey the concept of motion by using a strobe photograph, we do obtain a more accurate measure of the motion.

#### Figure 2

Strobe photograph showing the motion of a ball on the end of a string.





#### **Figure 3** Strobe photograph of a moving object. In this photograph, the time between flashes is so long that the motion is difficult to understand.

The photograph in Figure (2) was taken with the strobe flashing five times per second while the ball was moving slowly. As a result, we see a smooth curve and have a fairly complete idea of the ball's entire motion. When we run the strobe at a rate of five flashes per second but move the ball more rapidly in a complicated pattern, the result is as shown in Figure (3). From this picture it is difficult to guess the ball's path; thus Figure (3) provides us with a poor representation of the motion of the ball. But if we turn the strobe up from 5 to 15 flashes per second (as in Figure 4), the rapid and complicated motion of the ball is easily understood.

The motion of any object can be described by locating its position at successive intervals of time. A strobe photograph is particularly useful because it shows the position at equal time intervals throughout the picture; that is, in Figure (2) at intervals of 1/5 sec and in Figure (4) at intervals of 1/15 sec. For this text, we will use a special symbol,  $\Delta t$ , to represent the time interval between flashes of the strobe. The t stands for time, while the  $\Delta$  (Greek letter delta) indicates that these are short time intervals between flashes. Thus,  $\Delta t = 1/5$  sec in Figures (2) and (3), and  $\Delta t = 1/15$  sec in Figure (4).

For objects that are moving slowly along fairly smooth paths, we can use fairly long time intervals  $\Delta t$  between strobe flashes and their motion will be adequately described. As the motion becomes faster and more complicated, we turn the strobe up to a higher flashing rate to follow the object, as in Figure (4). To study complicated motion in more detail, we locate the position of the object after shorter and shorter time intervals  $\Delta t$ .



# **DISPLACEMENT VECTORS**

When we represent the motion of an object by a strobe photograph, we are in fact representing this motion by a series of displacements, the successive displacements of the object in equal intervals of time. Mathematically, we can describe these displacements by a series of displacement vectors, as shown in Figure (5). This illustration is a reproduction of Figure (2) with the successive displacement vectors drawn from the center of the images.

#### Figure 4

Strobe photograph of a similar motion. In this photograph, the time between flashes was reduced and the motion is more easily understood.

#### Figure 5

Displacement vectors. The displacement between flash number 1 and flash number 2 is represented by the displacement vector  $\vec{s}_1$  and so on. The entire path taken by the ball is represented by the series of eight displacement vectors.







(b)





(d)



Figure 6

Representation of the path of a ball for various  $\Delta t$ . As the shorter and shorter  $\Delta t$  is used, the path of the ball is more accurately represented, as in figures (b) through (d).

In a sense we are approximating the path of the ball by a series of straight lines along the path. This is reasonably accurate provided that  $\Delta t$  is short enough, as shown in Figure (6).

In Figure (6), (a) is the strobe photograph shown in Figure (4), taken at a strobe interval of  $\Delta t = 1/15$  sec; (b) shows how this photograph would have looked if we had set the strobe for  $\Delta t = 10/15$  sec, or 2/3 sec. Only one out of ten exposures would have been produced. If we had represented the path of the ball by the vector  $\vec{s}_1$  it would have been a gross misrepresentation. In (c), which would be the strobe picture at  $\Delta t = 6/15$  sec, we see that the ball is no longer moving in a straight line, but still  $\vec{s}_1$  and  $\vec{s}_2$ provide a poor representation of the true motion. Cutting  $\Delta t$  in half to get (d),  $\Delta t = 3/15$  sec, we would discover that there is a kink in the path of the ball. While taking the picture, we would have had to be careful in noticing the sequence of positions in order to draw the correct displacement vectors.

Reducing  $\Delta t$  to 2/15 sec (e), would give us a more detailed picture of the kink. This is not too different from (d); moreover, we begin to suspect that the

seven displacement vectors in (e) represent the path fairly accurately. When we reduce  $\Delta t$  to 1/15 sec (f), we get more pictures of the same kink and the curve becomes smoother. It now appears that in most places the 14 displacement vectors form a fairly accurate picture of the true path. We notice, however, that the very bottom of the kink is cut off abruptly; here, shorter time intervals are needed to get an accurate picture of the motion.

### A Coordinate System

In the strobe photographs discussed so far, we have a precise idea of the time scale, 1/5 second between flashes in Figure (2), 1/15 second in Figure (4), but no idea about the distance scale. As a result we know the direction of the succeeding displacement vectors, but do not know their magnitude.

One way to introduce a distance scale is to photograph the motion in front of a grid as shown in Figure (7). With this setup we obtain photographs like that shown in Figure (8), where we see the strobe motion of a steel ball projectile superimposed on the grid. The grid is illuminated by room lights which are dimmed to balance the exposure of the grid and the strobe flashes.



#### Figure 7

Experimental setup for taking strobe photographs. A Polaroid camera is used record the motion of a ball moving in front of a grid. The grid, made of stretched fish line, is mounted in front of a black painted wall.



**Figure 8** Strobe photograph of a steel ball projectile. The strobe flashes were 1/10 second apart.

Using techniques like that illustrated in Figure (9) to locate the centers of the images, we can transfer the information from the strobe photograph to graph paper and obtain the results shown in Figure (10). Figure (10) is the end result of a fair amount of tedious lab work, and the starting point for our analysis. For those who do not have strobe facilities, or the time to extract the information from a strobe photograph, we will include in the text a number of examples already transferred to graph paper in the form of Figure (10).

Using a television camera attached to an Apple II computer, we can, in under 2 minutes, obtain results that look like Figure (10). We will include a few of these *computer strobe photographs* in our examples of motion. However the computer strobe is not yet commercially available because we plan to use a computer with more modern graphics capabilities. It is likely that within a few years, one will be able to easily and quickly obtain results like those in Figure

(10). The grid, which has now become the graph paper in Figure (10), serves as our coordinate system for locating the images.

# **Manipulation of Vectors**

Figure (10) represents the kind of experimental data upon which we will base our description of motion. We have, up to now, described the motion of the projectile in terms of a series of displacement vectors labeled  $\vec{s}_{-1}$ ,  $\vec{s}_0$ ,  $\cdots \vec{s}_3$  as shown. To go further, to introduce concepts like velocity and acceleration, we need to perform certain routine operations on these displacement vectors, like adding and subtracting them. A number of vector operations were discussed in Chapter 2, let us briefly review here those that we need for the analysis of strobe photographs. We will also introduce the concept of a *coordinate vector* which will be useful in much of our work.



#### Figure 9

Using a pin and cylinder to locate the center of the ball. Move the cylinder until it just covers the image of the ball and then gently press down on the pin. The pin prick will give an accurate location for the center.



**Figure 11** Measuring the length of the vector  $\vec{S_1}$ .

# Measuring the Length of a Vector

One of the first pieces of information we need from a strobe photograph is the magnitude or length of the displacement vectors we have drawn. Figure (11) illustrates the practical way to obtain the lengths of the individual vectors from a graph like Figure (10). Take a piece of scrap paper and mark off the length of the vector as shown in the upper part of the figure. Then rotate the paper until it is parallel to the grid lines, and note the distance between the marks.



#### Figure 10

Strobe photograph transferred to graph paper. Using the pin and cylinder of Figure (9), we located the coordinates of the center of each image in Figure (8), and then reconstructed the strobe photograph as shown. We can now perform our analysis on the large graph paper rather than the small photograph.

In Figure (11), we see that the marks are 20 small grid spacings apart. In Figure (10), we see that each grid spacing represents a distance of 1 centimeter. Thus in Figure (11), the vector  $\vec{s}_1$  has a magnitude of 20 centimeters. We can write this formally as

 $|\vec{s}_1| = 20 \text{ cm}$ 

This technique may seem rather simple, but it works well and you will use it often.

# **Graphical Addition and Subtraction**

Since we are working with experimental data in graphical form, we need to use graphical techniques to add and subtract vectors. These techniques, originally introduced in Chapter 2, are reviewed here in Figures (12) and (13). Figure (12a) and (12b) show the addition of two vectors by placing them head to tail. Think of the vectors  $\vec{A}$  and  $\vec{B}$  as separate trips; the sum  $(\vec{A} + \vec{B})$  is our net displacement as we take the trips  $\vec{A}$  and  $\vec{B}$  in succession. To subtract  $\vec{B}$  from  $\vec{A}$ , we simply add  $(-\vec{B})$  to  $\vec{A}$  as shown in Figure (12c).

To perform vector addition and subtraction, we need to move the vectors from one place to another. This is easily done with a triangle and a straight edge as indicated in Figure (13). The triangle and straight edge allows you to draw a parallel line; then mark a piece of paper as in Figure (11), to make the new vector have the same length as the old one.



**Figure 12** Addition and subtraction of vectors.

For those who are mathematically inclined, this simple graphical work with vectors may seem elementary, especially compared to the exercises encountered in an introductory calculus course. But, as we shall see, this graphical work emphasizes the basic concepts. We will have many opportunities later to extract sophisticated formulas from these basic graphical operations.

For these exercises, you may use the practice graph on page 3-28, and the tear out sheet on page 3-29.

# Exercise 1

Find the magnitudes of the vectors  $\vec{s}_0$ ,  $\vec{s}_1$ ,  $\vec{s}_2$ , and  $\vec{s}_3$  in Figure (10).

# Exercise 2

Explain why the vector  $\vec{s}_{04}$ , given by

 $\vec{s}_{04} = \vec{s}_0 + \vec{s}_1 + \vec{s}_2 + \vec{s}_3$ 

has a magnitude of 91.3 cm which is quite a bit less than the sum of the lengths  $|s_0| + |s_1| + |s_2| + |s_3|$ .

# Exercise 3

Use graphical methods to find the vector  $(\vec{s}_3 - \vec{s}_2)$ . (The result should point vertically downward and have a length of about 10 cm.)



**Figure 13** *Moving vectors around. (This was discussed in Figure 2-12.)* 

# Coordinate System and Coordinate Vectors

A coordinate system allows us to convert graphical work into a numerical calculation that can, for example, be carried out on a computer. Figure (14) illustrates two convenient ways of describing the location of a point. One is to give the x and y coordinates of the point (x,y), and the other is to use a *coordinate vector*  $\vec{R}$  which we define as a vector that is drawn *from the origin of the coordinate system* to the point of interest.

Figure (15) illustrates the way an arbitrary vector  $\vec{S}$  can be expressed in terms of coordinate vectors. From the diagram we see that  $\vec{R}_2$  is the vector sum of  $\vec{R}_1 + \vec{S}$ , thus we can solve for the vector  $\vec{S}$  to get the result

$$\vec{S} = \vec{R}_2 - \vec{R}_1.$$



**Figure 14** The coordinate vector  $\vec{R}$ , which starts at the origin, locates the point (x,y).



Figure 15

Expressing the vector  $\vec{S}$  in terms of coordinate vectors.

# ANALYSIS OF STROBE PHOTOGRAPHS

In our analysis of the strobe photograph of projectile motion, Figure (10), we are representing the path of the ball by a series of displacement vectors  $\hat{S}_0 \dots \hat{S}_3$ (We will think of the photograph as starting at point (0). The point labeled (-1) will be used later in our calculation of the instantaneous velocity at point (0). In a sense, we "know" that the ball actually went along a smooth continuous curve, and we could have represented the curve more accurately by reducing  $\Delta t$  as we did in Figure (6). But with many images to mark the trajectory, each displacement vector  $S_i$ becomes too short for accurate graphical work. In taking a strobe photograph, one must reach a compromise where the displacement vectors  $\hat{S}_i$  are long enough to work with, but short enough to give a reasonable picture of the motion.

# Velocity

The series of displacement vectors in Figure (10) show not only the trajectory of the projectile, but because the images are located at equal time intervals, we also have an idea of the speed of the projectile along its path. A long displacement vector indicates a higher speed than a short one. For each of the displacement vectors we can calculate what one would call the average speed of the projectile during that interval.

The idea of an average speed for a trip should be fairly familiar. If, for example, you went on a trip for a total distance of 90 miles, and you took 2 hours, you divide 90 miles by 2 hours to get an average speed of 45 miles per hour. For more detailed information about your speed, you break the trip up into small segments. For example, if you wanted to know how fast you were moving down the interstate highway, you measure how long it takes to pass two consecutive mile markers. If it took one minute, then your average speed during this short time interval is one mile divided by 1/60 hour which is 60 miles per hour. If you broke the whole trip down into 1 minute intervals, measured how far you went during each interval, and calculated your average speed for each interval, you would have a fairly complete record of your speed during your trip. It is

this kind of record that we get from a strobe photograph of the motion of an object.

In physics, we use a concept that contains more information than simply the speed of the object. We want to know not only how many miles per hour or centimeters per second an object is moving, but also what direction the object is moving. This information is all contained in the concept of a *velocity vector*.

To construct a velocity vector for the projectile shown in Figure (10), when, for example, the ball is at position 1, we take the displacement vector  $\vec{S}_1$ , divide it by the strobe time interval  $\Delta t$ , to get what we will call the velocity vector  $\vec{v}_1$ :

$$\vec{v}_1 \equiv \frac{\vec{S}_1}{\Delta t} \tag{1}$$

In Equation (1), what we have done is multiply the vector  $\vec{S}_1$  by the number  $(1/\Delta t)$  to get  $\vec{v}_1$ . From our earlier discussion of vectors we know that multiplying a vector by a number gives us a vector that points in the same direction, but has a new length. Thus  $\vec{v}_1$  is a vector that points in the same direction as  $\vec{S}_1$ , but it now has a length given by

$$\left| \vec{\mathbf{v}}_{1} \right| = \frac{\left| \vec{\mathbf{S}}_{1} \right|}{\Delta t} = \frac{20 \text{ cm}}{.1 \text{ sec}} = 200 \frac{\text{cm}}{\text{sec}}$$

where we used  $|\vec{S}_1| = 20$  cm from Figure (11) and we knew that  $\Delta t = .1$  sec for this strobe photograph.

Not only have we changed the length of  $\vec{S}_1$  by multiplying by  $(1/\Delta t)$ , we have also changed the dimensions from that of a distance (cm) to that of a speed (cm/sec). Thus the velocity vector  $\vec{v}_1$  contains two important pieces of information. It points in the direction of the motion of the ball, and has a length or magnitude equal to the speed of the ball.

(Physics texts get rather picky over the use of the words speed and velocity. The word *speed* is reserved for the magnitude of the velocity, like 200 cm/sec. The word *velocity* is reserved for the velocity vector ity vector as defined above; the *velocity vector* 

describes both the direction of motion and the speed. We will also use this convention throughout the text.)

In our discussion of strobe photographs, we noted that if we used too long a time interval  $\Delta t$ , we got a poor description of the motion as in Figures (6b) and (6c). As we used shorter time intervals as in Figures (6d, e, and f), we got a better and better picture of the path.

We have the same problem in dealing with the velocity of an object. If we use a very long  $\Delta t$ , we get a crude, average, description of the object's velocity. As we use a shorter and shorter  $\Delta t$ , our description of the velocity, Equation (1), becomes more and more precise.

Since, in this chapter, we will be working with experimental data obtained from strobe photographs, there is a practical limit on how short a time interval  $\Delta t$  we can use and have vectors big enough to work with. We will see that, for the kinds of motion that we encounter in the introductory physics lab, a reasonably short  $\Delta t$  like .1 sec gives reasonably accurate results.

If you make more precise measurements of the position of an object you generally find that as you use shorter and shorter  $\Delta t$  to measure velocity, you reach a point where the velocity vector no longer changes. What happens is that you reach a point where, if you cut  $\Delta t$  in half, the particle goes in the same direction but only half as far. Thus both the displacement  $\vec{S}_1$  and the time interval  $\Delta t$  are both cut in half, and the ratio  $\vec{v}_1 = \vec{S}_1 / \Delta t$  is unchanged. This limiting process, where we see that the velocity vector changes less and less as  $\Delta t$  is reduced, is demonstrated graphically in our discussion of instantaneous velocity at the end of the chapter.

# **Exercise (4)**

What is the magnitude of the velocity vector  $\vec{v}_3$ , for the ball in Figure (10). Give your answer in cm/sec.

Equation (1) is well suited for graphical work but for numerical calculations it is convenient to express  $\vec{S}_i$  in terms of the coordinate vectors  $\vec{R}_i$ . This is done in Figure (16), where we see that the vector sum  $\vec{R}_i + \vec{S}_i = \vec{R}_{i+1}$  thus  $\vec{S}_i = \vec{R}_{i+1} - \vec{R}_i$  and Equation (1) becomes

$$\vec{v}_{i} = \frac{\left(\vec{R}_{i+1} - \vec{R}_{i}\right)}{\Delta t}$$
(2)

If we call  $(\vec{R}_{i+1} - \vec{R}_i)$  the "change in the position  $\vec{R}$  during the time  $\Delta t$ ", and denote this change by  $\Delta \vec{R}$ , Equation (2) becomes

$$\vec{v}_i = \frac{\left(\vec{R}_{i+1} - \vec{R}_i\right)}{\Delta t} = \frac{\Delta \vec{R}}{\Delta t}$$
 (2a)

which is perhaps a more familiar notation for those who have already studied calculus. In a calculus course, one would define the velocity  $\vec{v_i}$  by taking the limit as  $\Delta t \rightarrow 0$  (i.e., by turning the strobe flashing rate "all the way up"). In our experimental work with strobe photographs, we reduce  $\Delta t$  only to the point where we have a reasonable representation of the path; using too short a time interval makes the experimental analysis impossible.



Figure 16 Expressing the velocity vector  $\vec{v_i}$ , in terms of the coordinate vectors  $\vec{R_i}$  and  $\vec{R_{i+1}}$ .

#### Acceleration

In Chapter 1 on Einstein's special theory of relativity, we limited our discussion to uniform motion, motion in a straight line at constant speed. If we took a strobe photograph of an object undergoing uniform motion, we would get a result like that shown in Figure (17). All the velocity vectors would point in the same direction and have the same length. We will, from now on, call this *motion with constant velocity*, meaning that the velocity vector is constant, unchanging.



Motion with constant velocity.

From the principle of relativity we learned that there is something very special about motion with constant velocity—we cannot feel it. Recall that one statement of the principle of relativity was that there is no experiment that you can perform to detect your own uniform motion relative to empty space. You cannot tell, for example, whether the room you are sitting in is at rest or hurdling through space at a speed of 100,000 miles per hour.

Although we cannot feel or detect our own uniform motion, we can easily detect non uniform motion. We know what happens if we slam on the brakes and come to a sudden stop—everything in the car falls forward. A strobe photograph of a car using the brakes might look like that shown in Figure (18a). Each successive velocity vector gets shorter and shorter until the car comes to rest.



Figure 18a Put on the brakes, and your velocity changes.

Another way the velocity of a car can change is by going around a corner as illustrated in Figure (18b). In that figure the speed does not change, each velocity vector has the same length, but the directions are changing. It is also easy to detect this kind of change in velocity—all the packages in the back seat of your car slide to one side of the seat.



**Figure 18b** When you drive around a corner, your speed may not change, but your velocity vector changes in direction.

The point we want to get at is, *what do we feel when our velocity changes*? Consider two examples. In the first, we are moving at constant velocity, due east at 60 miles per hour. A strobe photograph showing our initial and final velocity vectors  $\vec{v}_i$  and  $\vec{v}_f$  would look like that in Figure (19a). If we define the *change in velocity*  $\Delta \vec{v}$  by the equation

$$\Delta \vec{v} \equiv \vec{v}_{f} - \vec{v}_{i}$$

then from Figure (19b) we see that  $\Delta \vec{v} = 0$  for uniform motion.

For the second example, suppose we are traveling due south at 60 miles per hour, and a while later are

(a) 
$$\overrightarrow{V_i}$$
  $\overrightarrow{V_f}$ 

(b) 
$$\overrightarrow{V_f}$$
  $\overrightarrow{\Delta V} = \overrightarrow{V_f} - \overrightarrow{V_i} = 0$ 

**Figure 19** We see that  $\overline{\Delta v} = 0$  for motion with constant velocity.

moving due east at 60 miles per hour, as indicated in Figure (20a). Now we have a non zero change in velocity  $\Delta \vec{v}$  as indicated in Figure (20b).

In our two examples, we find that if we have uniform motion which we cannot feel, the change in velocity  $\overrightarrow{\Delta v}$  is zero. If we have non uniform motion,  $\overrightarrow{\Delta v}$  is not zero and we can feel that. Is it  $\overrightarrow{\Delta v}$ , the change in velocity, that we feel?

Almost, but not quite. Let us look at our second example, Figure (20), more carefully. There are two distinct ways that our velocity can change from pointing south to pointing east. In one case there could have been a gradual curve in the road. It may have taken several minutes to go around the curve and we would be hardly aware of the turn.

In the other extreme, we may have been driving south, bounced off a stalled truck, and within a fraction of a second finding ourselves traveling due east. In both cases our change in velocity  $\overrightarrow{\Delta v} = \overrightarrow{v_f} - \overrightarrow{v_i}$  is the same, as shown in Figure (20b). But the effect on us is terribly different. The difference in the two cases is that the change in velocity  $\overrightarrow{\Delta v}$  occurred much more rapidly when we struck the truck than when we went around the curve. What we feel is not  $\overrightarrow{\Delta v}$  alone, but how fast  $\overrightarrow{\Delta v}$  happens.

If we take the change in velocity  $\overrightarrow{\Delta v}$  and divide it by the time  $\Delta t$  over which the change takes place, then the smaller  $\Delta t$ , the more rapidly the change takes place, the bigger the result. This ratio  $\overrightarrow{\Delta v}/\Delta t$  which more closely represents what we feel than  $\overrightarrow{\Delta v}$  alone, is given the special name *acceleration*.



**Figure 20**  $\overline{\Delta v} \neq 0$  when we change our direction of motion.

The physicists' use of the word acceleration for the quantity  $\overline{\Delta v}/\Delta t$  presents a problem for students. The difficulty is that we have grown up using the word *acceleration*, and already have some intuitive feeling for what that word means. Unfortunately this intuition usually does not match what physicists mean by acceleration. Perhaps physicists should have used a different name for  $\overline{\Delta v}/\Delta t$ , but this did not happen. The problem for the student is therefore not only to develop a new intuition for the quantity  $\overline{\Delta v}/\Delta t$ , but also to discard previous intuitive ideas of what acceleration might be. This can be uncomfortable.

The purpose of the remainder of this chapter is to develop a new intuition for the physics definition of acceleration. To do this we will consider three examples of motion; projectile motion, uniform circular motion, and projectile motion with air resistance. In each of these cases, which can be carefully studied in the introductory lab or simulated, we will use strobe photographs to determine how the acceleration vector  $\Delta v / \Delta t$  behaves. In each case we will see that there is a simple relationship between the behavior of the acceleration vector and the forces pulling or pushing on the object. This relationship between force and acceleration, which is the cornerstone of mechanics, will be discussed in a later chapter. Here our goal is to develop a clear picture of acceleration itself.

## Determining Acceleration from a Strobe Photograph

We will use strobe photographs to provide an explicit experimental definition of acceleration. In the next chapter we will see how the strobe definitions go over to the calculus definition that you may have already studied. We prefer to start with the strobe definition, not only because it provides a more intuitive approach to the concept, but also because of its experimental origin. With an experimental definition we avoid some conceptual problems inherent in calculus. It turns out, surprisingly, that some of the concepts involved in the calculus definition of acceleration are inconsistent with physics. We can more clearly understand these inconsistencies when we use an experimental definition of acceleration as the foundation for our discussion.

# The Acceleration Vector

The quantity  $\overline{\Delta v}/\Delta t$ , which we call acceleration, is usually denoted by the vector  $\vec{a}$ 

$$\vec{a} = \frac{\vec{\Delta v}}{\Delta t}$$
 (3)

where  $\Delta v$  is the change in the velocity vector during the time  $\Delta t$ . To see how to apply Equation (3) to a strobe photograph, suppose that Figure (21) represents a photograph of a particle moving with some kind of non uniform velocity. Labeling the image positions 1, 2, 3, etc. and the corresponding velocity vectors  $\vec{v}_1, \vec{v}_2, \vec{v}_3 \cdots$ , let us consider what the particle's acceleration was during the time it went from position 2 to position 3. At position 2 the particle's velocity was  $\vec{v}_2$ . When it got to position 3 its velocity was  $\vec{v}_3$ . The time it took for the velocity to change from  $\vec{v}_2$  to  $\vec{v}_3$ , a change  $\Delta \vec{v} = \vec{v}_3 - \vec{v}_2$ , was the strobe time  $\Delta t$ . Thus according to Equation (3), the particle's acceleration during the interval 2 to 3, which we will call  $\vec{a}_3$ , is given by

$$\vec{a}_3 \equiv \frac{\vec{\Delta v}_{23}}{\Delta t} = \frac{\vec{v}_3 - \vec{v}_2}{\Delta t}$$
 (4a)

(One could object to using the label  $\vec{a}_3$  for the acceleration during the interval 2 to 3. But a closer inspection shows that  $\vec{a}_3$  is an accurate name. Actually the velocity  $\vec{v}_3$  is the average velocity in the interval 3 to 4, and  $\vec{v}_2$  is the average velocity in the interval 2 to 3. Thus  $\Delta \vec{v} = \vec{v}_3 - \vec{v}_2$  is a change in velocity centered on position 3. As a result Equation (4a) gives surprisingly accurate results when working with experimental strobe photographs. In any case such errors become vanishingly small when we use sufficiently short  $\Delta t$ 's.)



Determining a for non uniform motion.

If we have a strobe photograph with many images, then by extending Equation (4a), the acceleration at position i is

$$\vec{a}_{i} \equiv \frac{\vec{\Delta v}_{i}}{\Delta t} = \frac{\vec{v}_{i} - \vec{v}_{i-1}}{\Delta t}$$
 strobe definition of acceleration (4)

We will call Equation (4) our *strobe definition of acceleration*. Implicit in this definition is that we use a short enough  $\Delta t$  so that all the kinks in the motion are visible, but a long enough  $\Delta t$  so that we have vectors long enough to work with.



#### Figure 10a

A section of the projectile motion photograph, Figure (10), showing the displacement vectors  $\vec{S}_1$  and  $\vec{S}_2$ .



The vector  $\vec{S}_2 - \vec{S}_1$  points straight down and has a length of about 10 cm.

# **PROJECTILE MOTION**

As our first example in the use of our strobe definition of acceleration, let us calculate the acceleration of the ball at position 2 in our strobe photograph, Figure (10), of projectile motion.

The first problem we face is that Equation (4) expresses the acceleration vector  $\vec{a}_2$  in terms of the velocity vectors  $\vec{v}_1$  and  $\vec{v}_2$ , while the strobe photograph shows only the displacement vectors  $\vec{S}_1$  and  $\vec{S}_2$ , as seen in Figure (10a), a segment of Figure (10) reproduced here.

The easiest way to handle this problem is to use the formulas

$$\vec{v}_1 = \frac{\vec{S}_1}{\Delta t}; \qquad \vec{v}_2 = \frac{\vec{S}_2}{\Delta t}$$

in Equation (4a) to express  $\vec{a}_2$  directly in terms of the known vectors  $\vec{S}_1$  and  $\vec{S}_2$ . The result is

$$\vec{a}_{2} = \frac{\vec{v}_{2} - \vec{v}_{1}}{\Delta t} = \frac{(\vec{S}_{1}/\Delta t) - (\vec{S}_{2}/\Delta t)}{\Delta t}$$
$$\vec{a}_{2} = \frac{\vec{S}_{2} - \vec{S}_{1}}{\Delta t^{2}} \qquad experimental measurment of acceleration$$
(5)

Equation (5) tells us that we can calculate the acceleration vector  $\vec{a}_2$  by first constructing the vector  $\vec{S}_2 - \vec{S}_1$ , and then dividing by  $\Delta t^2$ . That means that  $\vec{a}_2$  points in the direction of the vector  $\vec{S}_2 - \vec{S}_1$ , and has a length equal to the length  $|\vec{S}_2 - \vec{S}_1|$  (in cm) divided by  $\Delta t^2$ . As a result the magnitude of the acceleration vector has the dimensions of cm/sec<sup>2</sup>.

Let us apply Equation (5) to our projectile motion photograph, Figure (10), to see how all this works. The first step is to use *vector subtraction* to construct the *vector*  $\vec{S}_2 - \vec{S}_1$ . This is done in Figure (22). First we draw the vectors  $\vec{S}_1$  and  $\vec{S}_2$ , and then construct the vector  $-\vec{S}_1$  as shown. (The vector  $-\vec{S}_1$  is the same as  $\vec{S}_1$  except that it points in the opposite direction.) Then we add the vectors  $\vec{S}_2$  and  $-\vec{S}_1$  to get the vector  $\vec{S}_2 - \vec{S}_1$  by the usual technique of vector addition as shown

 $\vec{S}_2 - \vec{S}_1 = \vec{S}_2 + (-\vec{S}_1)$  (6)

Note that even if  $\vec{S}_2$  and  $\vec{S}_1$  had the same length, the difference  $\vec{S}_2 - \vec{S}_1$  would not necessarily be zero because this is vector subtraction, *NOT NUMERICAL* SUBTRACTION.

Once we have constructed the vector  $\vec{S}_2 - \vec{S}_1$ , we know the direction of the acceleration vector  $\vec{a}_2$  because it points in the *same direction* as  $\vec{S}_2 - \vec{S}_1$ . In Figure (22), we see that  $\vec{S}_2 - \vec{S}_1$  points straight down, thus  $\vec{a}_2$  points straight down also.

Now that we have the direction of  $\vec{a}_2$ , all that is left is to calculate its magnitude or length. This magnitude is given by the formula

$$\vec{a}_2 = \frac{\left|\vec{S}_2 - \vec{S}_1\right|}{\Delta t^2} = \frac{\text{length of vector } \vec{S}_2 - \vec{S}_1}{\Delta t^2}$$
(7)

To get the length of the  $\vec{S}_2 - \vec{S}_1$ , we can use the technique shown in Figure (11). Mark off the length of the vector  $\vec{S}_2 - \vec{S}_1$  on a piece of scrap paper, and then use the grid to see how many centimeters apart the marks are. In this case, where  $\vec{S}_2 - \vec{S}_1$  points straight down, we immediately see that  $\vec{S}_2 - \vec{S}_1$  is about 10 cm long. Thus the magnitude of  $\vec{a}_2$  is given by

$$\left|\vec{a}_{2}\right| = \frac{\left|\vec{S}_{2} - \vec{S}_{1}\right|}{\Delta t^{2}} = \frac{10 \text{ cm}}{(.1 \text{ sec})^{2}} = 1000 \frac{\text{cm}}{\text{sec}^{2}}$$
 (8)

where we knew  $\Delta t = .1$  sec for the strobe photograph in Figure (10).

Our conclusion is that, at position 2 in the projectile motion photograph, the ball had an acceleration  $\vec{a}_2$  that pointed straight down, and had a magnitude of about 1000 cm/sec<sup>2</sup>.

For this exercise, you may use the tear out sheet on page 3-30.

#### **Exercise** 5

(**Do this now before reading on**.) Find the acceleration vectors  $\vec{a}_0, \vec{a}_1$ , and  $\vec{a}_3$  for the projectile motion in Figure (10). From your results, what can you say about the acceleration of a projectile?

To give the reader some time to think about the above exercise on projectile motion, we will change the topic for a while and analyze what is called *uniform circular motion*. In uniform circular motion, the particle travels like a speck of dust sitting on a revolving turntable.

The explicit example we would like to consider is a golf ball with a string attached, being swung in a circle over the instructor's head, as indicated in Figure (23a). We could photograph this motion, but it is very easy to simulate a strobe photograph of uniform circular motion by drawing a circle with a compass, and marking off equal intervals as shown in Figure (23b). In that figure we have also sketched in the displacement vectors as we did in our analysis of the projectile motion photograph.







#### Figure 23b

Simulating a strobe photograph of a golf ball swinging at constant speed in a circle. We marked off equal distances using a compass. Figure (23b) shows the kind of errors we have to deal with in using a strobe to study motion. Clearly the golf ball travels along the smooth circular path rather than the straight line segments marked by the vectors . As we use shorter and shorter  $\Delta t$  our approximation of the path gets better and better, but soon the vectors get too short for accurate graphical work. Choosing images spaced as in Figure (23b) gives vectors a reasonable length, and a reasonable approximation of the circular path.

(It will turn out that when we use our strobe definition of acceleration, most errors caused by using a finite  $\Delta t$  cancel, and we get a very accurate answer. Thus we do not have to worry much about how far apart we draw the images.)

Now that we have the displacement vectors we can construct the acceleration vectors  $\vec{a}_1, \vec{a}_2, \cdots$  using Equation (5). The construction for  $\vec{a}_2$  is shown in Figure (24). To the vector  $\vec{S}_2$  we add the vector  $-\vec{S}_1$ to get the vector  $\vec{S}_2 - \vec{S}_1$  as shown.

The first thing we note is that the vector  $\vec{S}_2 - \vec{S}_1$ points toward the center of the circle! Thus the acceleration vector  $\vec{a}_2$  given by

$$\vec{a}_2 = \frac{\vec{S}_2 - \vec{S}_1}{\Delta t^2} \tag{9}$$

also points toward the center of the circle.



We find that the vector  $\vec{S}_2 - \vec{S}_1$ , and therefore the acceleration, points toward the center of the circle.

# Exercise (6)

(Do this now.) Find the direction of at least 4 more acceleration vectors around the circle. In each case show that  $\vec{a}_i$  points toward the center of the circle.

We said earlier that the physicists' definition of acceleration, which becomes  $\vec{a}_i = (\vec{S}_i - \vec{S}_{i-1})/\Delta t^2$ , does not necessarily agree with your own intuitive idea of acceleration. We have just discovered that, using the physicists' definition, a particle moving at constant speed along a circular path *accelerates* toward the center of the circle. Unless you had a previous physics course, you would be unlikely to guess this result. It may seem counter intuitive. But, as we said, we are using these examples to develop an intuition for the physics definition of acceleration. Whether you like it or not, according to the physics definition, a particle moving at constant speed around a circle, is accelerating toward the center. In a little while, the reason for this will become clear.

# Magnitude of the Acceleration for Circular Motion

Although perhaps not intuitive, we have gotten a fairly simple result for the direction of the acceleration vector for uniform circular motion. The center is the only unique point for a circle, and that is where the acceleration vector points. The next thing we need to know is how long the acceleration vectors are; what is the magnitude of this center pointing acceleration. From the strobe definition, the magnitude  $|\vec{a}_2|$  is  $\vec{a}_2 = |\vec{S}_2 - \vec{S}_1|/\Delta t^2$ , a rather awkward result that appears to depend upon the size of  $\Delta t$  that we choose. However with a bit of geometrical construction we can re-express this result in terms of the particle's speed v and the circle's radius r. The derivation is messy, but the result is simple. This is one case, where, when we finish the derivation, we recommend that the student memorize the answer rather than try to remember the derivation. Uniform circular motion appears in a number of important physics problems, thus the formula for the magnitude of the acceleration is important to know.

In Figure (25a) we have constructed two triangles, which are shown separately in Figures (25b) and (25c). As seen in Figure (25b), the big triangle which goes from the center of the circle to positions (1) and (2) has two equal sides of length r, the radius of the circle, and one side whose length is equal to the particle's speed v times the strobe time  $\Delta t$ .

The second triangle, shown in Figure (25c), has sides of length  $|\vec{S}_2|$  and  $|-\vec{S}_1|$ , but both of these are of length v $\Delta t$  as shown. The third side is of length  $|\vec{S}_2 - \vec{S}_1|$ , the length we need for our calculation of the magnitude of the acceleration vector.

The trick of this calculation is to note that the angles labeled  $\theta$  in Figures (29b, c) are the same angle, so that these two triangles are similar isosceles triangles. The proof that these angles are equal is given in Figure (26) and its caption.

With similar triangles we can use the fact that the ratios of corresponding sides are equal. Equating the ratio of the short side to the long side of the triangle of Figure (25b), to the ratio of the short side to the long side of the triangle in Figure (25c), we get



#### Figure 25

Derivation of the formula for the magnitude of the acceleration of a particle with uniform circular motion.

$$\frac{v\Delta t}{r} = \frac{\left|\vec{S}_2 - \vec{S}_1\right|}{v\Delta t} \tag{10}$$

Multiplying Equation (10) through by v and dividing both sides by  $\Delta t$  gives

$$\frac{v^2}{r} = \frac{\left|\vec{S}_2 - \vec{S}_1\right|}{\Delta t^2} = \left|\vec{a}_1\right|$$
(11)

we got  $|\vec{s}_2 - \vec{s}_1| / \Delta t^2$  on the right side, but this is just the magnitude of the acceleration vector  $\vec{a}_1$ . Since the same derivation applies to any position around the circle, we get the simple and general result that, for a particle moving with uniform circular motion, the particle's acceleration  $\vec{a}$  points toward the center of the circle, and has a magnitude



where v is the speed of the particle and r is the radius of the circle. As we said, this simple result should be memorized.



#### Figure 26

That the two angles labeled  $\theta$  in Figure (25) are the same, may be seen in the following geometrical construction. Since the sum of the angles in any triangle is 180°, we get  $\alpha + \varphi + 90^\circ = 180^\circ$  (from triangle BCD). Because BAC is an isosceles triangle,  $(90^\circ - \theta) + \alpha = \varphi$ . Eliminating  $\varphi$  we get  $2\alpha = \theta$ , which is the result we expected.

# AN INTUITIVE DISCUSSION OF ACCELERATION

We have now studied two examples of non uniform motion, the projectile motion seen in the strobe photograph of Figure (10), and the circular motion of a golf ball on the end of a string, a motion we illustrated in Figure (23). In each case we calculated the acceleration vector of the particle at different points along the trajectory. Let us now review our results to see if we can gain some understanding of why the acceleration vector behaves the way it does.

If you worked Exercise (5) correctly, you discovered that all the acceleration vectors are the same, at least to within experimental accuracy. As shown in Figure (27), as the steel ball moves along its trajectory, its acceleration vector points downward toward the earth, and has a constant magnitude of about 1000 cm/sec<sup>2</sup>.

As shown in Figure (28), the golf ball being swung at constant speed around in a circle on the end of a string, accelerates toward the center of the circle, in the direction of the string pulling on the ball. The magnitude of the acceleration has the constant value.



All the acceleration vectors for projectile motion point down toward the earth.

We said that the string was pulling on the ball. To see that this is true, try swinging a ball on the end of a string (or a shoe on the end of a shoelace) in a circle. To keep the ball (or shoe) moving in a circle, you have to pull in on the string. In turn, the string pulls in on the ball (or shoe). If you no longer pull in on the string, i.e., let go, the ball or shoe flies away and no longer undergoes circular motion. The string pulling on the ball is necessary in order to have circular motion.

What is the common feature of projectile and circular motion? In both cases the object accelerates in the direction of the force acting on the object. When you throw a steel ball in the air, the ball does not escape earth's gravity. As the ball moves through the air, gravity is constantly pulling down on the ball. The result of this gravitational pull or force is to accelerate the ball in the direction of the gravitational force. That is why the projectile motion acceleration vectors point down toward the earth.

When we throw a ball a few feet up in the air, it does not get very far away from the surface of the earth. In other words we expect the gravitational pull to be equally strong throughout the trajectory. If the ball's acceleration is related to the gravitational pull, then we expect the acceleration to also be constant throughout the trajectory. Thus it is not surprising that all the vectors have the same length in Figure (27).



**Figure 28** The golf ball accelerates in the direction of the string that is pulling on it.

In the case of circular motion, the string has to pull in on the golf ball to keep the ball moving in a circle. As a result of this pull of the string toward the center of the circle, the ball accelerates toward the center of the circle. Again the acceleration is in the direction of the force on the object.

This relationship between force and acceleration, which we are just beginning to see in these two examples, forms the cornerstone of what is called *classical* or *Newtonian mechanics*. We have more details to work out, but we have just glimpsed the basic idea of much of the first half of this course. To give historical credit for these ideas, it was Galileo who first saw the importance of the concept of acceleration that we have been discussing, and Isaac Newton who pinned down the relationship between force and acceleration.

#### Acceleration Due to Gravity

Two more topics, both related to projectile motion, will finish our discussion in this section. The first is the fact that, if we can neglect air resistance, all projectiles near the surface of the earth have the same downward acceleration  $\vec{a}$ . If a steel ball and a feather are dropped in a vacuum, they fall together with the same acceleration. This acceleration, which is caused by gravity, is called the *acceleration due to gravity* and is denoted by the symbol  $\vec{g}$ . The vector  $\vec{g}$  points down toward the earth, and, at the surface of the earth, has a magnitude.

	acceleration due to	
$\left  \overrightarrow{g} \right  \equiv g = 980 \text{ cm/sec}^2$	gravity at the sur- face of the earth	(13)

This is quite consistent with our experimental result of about  $1000 \text{ cm/sec}^2$  that we got from the analysis of the strobe photograph in Figure (10).

If we go up away from the earth, the acceleration due to gravity decreases. At an altitude of 1,600 miles, the acceleration is down to half its value, about 500 cm/sec<sup>2</sup>. On other planets g has different values. For example, on the moon, g is only about 1/6 as strong as it is here on the surface of the earth, i.e.

$$g_{moon} = 167 \text{ cm/sec}^2 \tag{14}$$

From the relationship we have seen between force and acceleration we can understand why a projectile that goes only a few feet above the surface of the earth should have a constant acceleration. The gravitational force does not change much in those few feet, and therefore we would not expect the acceleration caused by gravity to change much either.

On the other hand there is no obvious reason, at this point, why in the absence of air, a steel ball and a feather should have the same acceleration. Galileo believed that all projectiles, in the absence of air resistance, have the same acceleration. But it was not until Newton discovered both the laws of mechanics (the relationship between acceleration and force) and the law of gravity, that it became a physical prediction that all projectiles have the same gravitational acceleration.

In the early part of the 20th century, Einstein went a step farther than Newton, and used the fact that all objects have the same gravitational acceleration to develop a geometrical interpretation of the theory of gravity. The gravitational force was reinterpreted as a curvature of space, with the natural consequence that a curvature of space affects all objects in the same way. This theory of gravity, known as *Einstein's general theory of relativity*, was a result of Einstein's effort to make the theory of gravity consistent with the principle of relativity.

It is interesting how the simplest ideas, the principle of relativity, and the observation that the gravitational acceleration is the same for all objects, are the cornerstones of one of the most sophisticated theories in physics, in this case Einstein's general theory of relativity. Even today, over three quarters of a century since Einstein developed the theory, we still do not understand what many of the predictions or consequences of Einstein's theory will be. It is exciting, for these predictions may help us understand the behavior of the universe from its very beginning.

#### Exercise (7)

The first earth satellite, Sputnik 1, traveled in a low, nearly perfect, circular orbit around the earth as illustrated in Figure (29).

(a) What was the direction of Sputnik 1's acceleration vector as it went around the earth?

(b) What was the direction of the force of gravity on Sputnik 1 as the satellite went around the earth?

(c) How is this problem related to the problem of the motion of the golf ball on the end of a string? Give an answer that your roommate, who has not had a physics course, would understand.



Sputnik 1's circular orbit.

# Projectile Motion with Air Resistance

Back to a more mundane subject, we wish to end this discussion of acceleration with the example of projectile motion with air resistance. Most introductory physics texts avoid this topic because they cannot deal with it effectively. Using calculus, one can handle only the simplest, most idealized examples, and even then the analysis is beyond the scope of most texts. But using strobe photographs it is easy to analyze projectile motion with air resistance, and we learn quite a bit from the results.

What turned out to be difficult, was to find an example where air resistance affected the motion of a projectile enough to produce a noticeable effect. We found that a golf ball and a ping pong ball have almost the same acceleration when thrown in the air, despite the considerable difference in weight or mass. Only when we used the rough surfaced Styrofoam balls used for Christmas tree ornaments did we finally get enough air resistance to give a significant effect.



**Figure 30a** Motion of a Styrofoam ball. This is the lightest ball we could find.

A strobe photograph of the projectile motion of the Styrofoam ball is seen in Figure (30a), and an analysis showing the resulting acceleration vectors in Figure (30b). In Figure (30b) we have also drawn the acceleration vectors  $\vec{g}$  that the ball would have had if there had not been any air resistance. We see that the effect of air resistance is to bend back and shorten the acceleration vectors.

Figure (31) is a detailed analysis of the Styrofoam's acceleration at point (3). (We used an enlargement of the strobe photograph to improve the accuracy of our work, such detailed analysis is difficult using small Polaroid photographs.) In Figure (31) $\vec{v}_3$  is the velocity of the ball,  $\vec{g}$  is the acceleration due to gravity, and  $\vec{a}_3$  the ball's actual acceleration. The vector  $\vec{a}_{air}$ , which represents the change in  $\vec{a}$  caused by air resistance is given by the vector equation

$$\vec{a}_3 = \vec{g} + \vec{a}_{air} \tag{15}$$

The important feature of Figure (31) is that  $\vec{a}_{air}$  is oppositely directed to the ball's velocity  $\vec{v}_3$ . To understand why, imagine that you are the stick



What we learn from this example is that <u>if we have</u> two forces simultaneously acting on an object, each force independently produces an acceleration, and the net acceleration is the vector sum of the independent accelerations. In this case the independent accelerations are caused by gravity and the wind. The net acceleration  $\vec{a}_3$  of the ball is given by the vector Equation (15),  $\vec{a}_3 = \vec{g} + \vec{a}_{air}$ . As we will see in later chapters, this vector addition of accelerations plays a fundamental role in mechanics.



#### Figure 31

When we do a detailed comparison of  $\vec{a}$  and  $\vec{g}$  at point 3, we see that the air resistance produces an acceleration  $\vec{a}_{aii}$  that points in the direction of the wind felt by the ball.

#### **Figure 30b** Acceleration of the Styrofoam ball.

# INSTANTANEOUS VELOCITY

In calculus, instantaneous velocity is defined by starting with the equation  $\vec{v}_i = (\vec{R}_{i+1} - \vec{R}_i)/\Delta t$  and then taking the limiting value of  $\vec{v}_i$  as we use shorter and shorter time steps  $\Delta t$ . This corresponds in a strobe photograph to using a higher and higher flashing rate which would give increasingly short displacement vectors  $\vec{S}_i$ . In the end result one pictures the instantaneous velocity being defined at each point along the continuous trajectory of the object.

The effect of using shorter and shorter  $\Delta t$  is illustrated in Figure (32). In each of these sketches the dotted line represents the smooth continuous trajec-

tory of the ball. In Figure (32a) where  $\Delta t = 0.4$  sec and there are only two images the only possible definition of  $\vec{v}_0$  is the displacement between these images, divided by  $\Delta t$  as shown. Clearly  $\Delta t$  is too large here for an accurate representation of the ball's motion.

A better description of motion is obtained in Figure (32b) where  $\Delta t = 0.1$  sec as in the original photograph. We used this value of in our analysis of the projectile motion, Figure (10). Reducing  $\Delta t$  by another factor 1/4 gives the results shown in Figure (32c). At this point the images provide a detailed picture of the path and  $\vec{v}_0 = \vec{S}_0/\Delta t$  is now tangent to the path at (0). A further decrease in  $\Delta t$  would produce a negligible change in  $\vec{v}_0$ .



The instantaneous velocity at point (0) is the final value of  $\vec{v}_0$ , the value illustrated in Figure (32d) which no longer changes as  $\Delta t$  is reduced. This is an abstract concept in that we are assuming such a final value exists. We are assuming that we always reach a point where using a stroboscope with a still higher flashing rate produces no observable change in the value of  $\vec{v}_0$ . This assumption, which has worked quite well in the analysis of large objects such as ping pong balls and planets, has proven to be false when investigated on an atomic scale. According to the quantum theory which replaces classical mechanics on an atomic scale when one uses a sufficiently short  $\Delta t$  in an attempt to measure velocity, the measurement destroys the experiment rather than giving a better value of  $\vec{v}_0$ .

# Instantaneous Velocity from a Strobe Photograph

In the case of projectile motion (i.e., motion with constant acceleration) there is a simple yet precise method for determining an object's instantaneous velocity  $\vec{v}_i$  from a strobe photograph. (Vectors representing instantaneous velocity will be underlined in order to distinguish them from the vectors representing the strobe definition of velocity.) This method, which also gives quite good approximate values for other kinds of motion, will be used in our computer calculations for determining the initial velocity of the object.

To see what the method is, consider Figure (33) where we have drawn the vector obtained from Figure (32d). We have also drawn a line from the center of image (-1) to the center of image (+1) and notice that  $\underline{\vec{v}}_i \Delta t$  is parallel to and precisely half as long as this line. Thus we can construct  $\underline{\vec{v}}_i \Delta t$  by connecting the preceding and following images and taking half of that line.

The vector constructed by the above rule is actually the average of the preceding velocity vector  $\vec{v}_{-1}$  and the following vector  $\vec{v}_0$ .

$$\vec{\mathbf{v}}_{i} = \frac{\vec{\mathbf{v}}_{-1} + \vec{\mathbf{v}}_{0}}{2} \tag{16}$$

as illustrated in Figure (33). (Note that the vector sum  $(\vec{v}_{-1} + \vec{v}_0)\Delta t$  is the same as the line  $2\vec{v}_i\Delta t$  which connects the preceding and following image.) This is a reasonable estimate of the ball's instantaneous velocity because  $\vec{v}_{-1}$  is the average velocity during the time  $\Delta t$  before the ball got to (0), and  $\vec{v}_0$  the average velocity during the interval after leaving (0). The ball's velocity at (0) should have a value intermediate between  $\vec{v}_{-1}$  and  $\vec{v}_0$ , which is what Equation (16) says.

The constant acceleration formula

$$\vec{S} = \underline{\vec{v}}_i t + \frac{1}{2} \vec{a} t^2 \tag{17}$$

which may be familiar from a high school physics course, provides a direct application of the concept of instantaneous velocity. (Remember that this is not a general formula; it applies only to motion with constant acceleration where the vector  $\vec{a}$  changes neither in magnitude or direction.) As illustrated in Figure (34) the total displacement of the projectile





**Figure 34** *Illustration of the constant acceleration formula as a vector equation.* 



during a time t (here t =  $3\Delta t$ ) is the vector sum of  $\vec{v_i} t$ and  $1/2\vec{a}t^2$ . To draw this figure, we used

$$\underline{\vec{v}}_{i}t = \underline{\vec{v}}_{i}(3\Delta t) = 3(\underline{\vec{v}}_{i}\Delta t)$$

and obtained  $\underline{\vec{v}}_i \Delta t$  from our method of determining instantaneous velocity. We also used

$$\frac{1}{2}\vec{a}t^2 = \frac{1}{2}\vec{a}(3\Delta t)^2 = \frac{9}{2}\vec{a}\Delta t^2$$

where we obtained  $\vec{a} \Delta t^2$  from the relation

$$\vec{a} = \frac{\vec{S}_2 - \vec{S}_1}{\Delta t^2}$$
$$\vec{a}\Delta t^2 = \left(\vec{S}_2 - \vec{S}_1\right)$$

as illustrated in Figure (35).



#### Figure 35

How to construct the vectors  $\vec{v}_i t$  and  $1/2\vec{a}t^2$  from a strobe photograph.

# For these exercises, use the tear out sheet on pages 3-31,32.

#### Exercise 8

Use Equation 17 to predict the displacements of the ball

- (a) Starting at position (0) for a total time  $t = 4\Delta t$ .
- (b) Starting at position (1) for a total time  $t = 3\Delta t$ .

Do the work graphically as we did in Figures 33-35.

### Exercise 9

The other constant acceleration formula is

$$\vec{\underline{v}}_{f} = \vec{\underline{v}}_{i} + \vec{a}t$$

where  $\vec{v}_i$  is the initial velocity, and  $\vec{v}_f$  the object's velocity a time t later. Apply this equation to Figure 10 to predict the ball's instantaneous velocity  $\vec{v}_f$  at point (3) for a ball starting at point (0). Check your prediction by graphically determining the instantaneous velocity at point (3).

Show your results on graph paper.

# Exercise 10

Show that the constant acceleration formulas would correctly predict projectile motion even if time ran backward. (For example, assume that the ball went backward as shown in Figure (36), and repeat Exercise 8b, going from position 3 to position 0.)



#### Figure 36

Run the motion of the ball backward in time, and it looks like it was launched from the lower right.













This is not a tear out page.





# Figure 1

Transition to instantaneous velocity.

# Chapter 4 Calculus in Physics

This chapter, which discusses the use of calculus in physics, is for those who have had a calculus course which they remember fairly well. For those whose calculus is weak or poorly remembered, or for those who have not studied calculus, you should replace this chapter with Chapter 1 of *Calculus 2000.* 

In the previous chapter we used strobe photographs to define velocity and acceleration vectors. The basic approach was to turn up the strobe flashing rate as we did in going from Figure (3-3) to (3-4) until all the kinks are clearly visible and the successive displacement vectors give a reasonable description of the motion. We did not turn the flashing rate too high, for the practical reason that the displacement vectors became too short for accurate work. Calculus corresponds to conceptually turning the strobe all the way up.

# LIMITING PROCESS

In our discussion of instantaneous velocity we conceptually turned the strobe all the way up as illustrated in Figures (2-32a) through (2-32d), redrawn here in Figure (1). In these figures, we initially see a fairly large change in  $\vec{v}_0$  as the strobe rate is increased and  $\Delta t$ reduced. But the change becomes smaller and it looks as if we are approaching some final value of  $\vec{v}_0$  that does not depend on the size of  $\Delta t$ , provided  $\Delta t$  is small enough. It looks as if we have come close to the final value in Figure (1c).

The progression seen in Figure (1) is called a *limiting process*. The idea is that there really is some true value of  $\vec{v}_0$  which we have called the instantaneous velocity, and that we approach this true value for sufficiently small values of  $\Delta t$ . This is a calculus concept, and in the language of calculus, we are *taking the limit as*  $\Delta t$  *goes to zero*.

# **The Uncertainty Principle**

For over 200 years, from the invention of calculus by Newton and Leibnitz until 1924, the limiting process and the resulting concept of instantaneous velocity was one of the cornerstones of physics. Then in 1924 Werner Heisenberg discovered what he called the *uncertainty principle* which places a limit on the accuracy of experimental measurements. Heisenberg discovered something very new and unexpected. He found that the act of making an experimental measurement unavoidably affects the results of an experiment. This had not been known previously because the effect on large objects like golf balls is undetectable. But on an atomic scale where we study small systems like electrons moving inside an atom, the effect is not only observable, it can dominate our study of the system.

One particular consequence of the uncertainly principle is that the more accurately we measure the position of an object, the more we disturb the motion of the object. This has an immediate impact on the concept of instantaneous velocity. If we turn the strobe all the way up, reduce  $\Delta t$  to zero, we are in effect trying to measure the position of the object with infinite precision. The consequence would be an infinitely big disturbance of the motion of the object we are studying. If we actually could turn the strobe all the way up, we would destroy the object we were trying to study.

It turns out that the uncertainty principle can have a significant impact on a larger scale of distance than the atomic scale. Suppose, for example, that we constructed a chamber 1 cm on a side, and wished to study the projectile motion of an electron inside. Using Galileo's idea that objects of different mass fall at the same rate, we would expect that the motion of the electron projectile should be the same as more massive objects. If we took a strobe photograph of the electron's



**Figure 2** *Hypothetical electron projectile motion experiment.* 

motion, we would expect get results like those shown in Figure (2). This figure represents projectile motion with an acceleration  $g = 980 \text{ cm/sec}^2$  and  $\Delta t = .01 \text{sec}$ , as the reader can easily check.

When we study the uncertainty principle in Chapter 30, we will see that a measurement that is accurate enough to show that Position (2) is below Position (1), could disturb the electron enough to reverse its direction of motion. The next position measurement could find the electron over where we drew Position (3), or back where we drew Position (0), or anywhere in the region in between. As a result we could not even determine what direction the electron is moving. This uncertainty would not be the result of a sloppy experiment, it is the best we can do with the most accurate and delicate measurements possible.

The uncertainty principle has had a significant impact on the way physicists think about motion. Because we now know that the measuring process affects the results of the measurement, we see that it is essential to provide experimental definitions to any physical quantity we wish to study. A conceptual definition, like turning the strobe all the way up to define instantaneous velocity, can lead to fundamental inconsistencies.

Even an experimental definition like our strobe definition of velocity can lead to inconsistent results when applied to something like the electron in Figure (2). But these inconsistencies are real. Their existence is telling us that the very concept of velocity is beginning to lose meaning for these small objects.

On the other hand the idea of the limiting process and instantaneous velocity is very convenient when applied to larger objects where the effects of the uncertainty principle are not detectable. In this case we can apply all the mathematical tools of calculus developed over the past 250 years. The status of instantaneous velocity has changed from a basic concept to a useful mathematical tool. Those problems for which this mathematical tool works are called problems in *classical physics*; and those problems for which the uncertainty principle is important, are in the realm of what we call *quantum physics*.
## CALCULUS DEFINITION OF VELOCITY

With the above perspective on the physical limitations on the limiting process, we can now return to the main topic of this chapter—the use of calculus in defining and working with velocity and acceleration.

In discussing the limiting process in calculus, one traditionally uses a special set of symbols which we can understand if we adopt the notation shown in Figure (3). In that figure we have drawn the coordinate vectors  $\vec{R}_i$  and  $\vec{R}_{i+1}$  for the i th and (i + 1) th positions of the object. We are now using the symbol  $\Delta \vec{R}_i$  to represent the displacement of the ball during the i to i+1 interval. The vector equation for  $\Delta \vec{R}_i$  is

$$\vec{\Delta R}_{i} = \vec{R}_{i+1} - \vec{R}_{i} \tag{1}$$

In words, Equation (1) tells us that  $\overrightarrow{\Delta R}_i$  is the change, during the time  $\Delta t$ , of the position vector  $\vec{R}$  describing the location of the ball.



**Figure 3** Definitions of  $\Delta \vec{R}_i$  and  $\vec{V}_i$ .

The velocity vector  $\vec{v}_i$  is now given by

$$\vec{v}_i = \frac{\vec{\Delta R}_i}{\Delta t} \tag{2}$$

This is just our old strobe definition  $\vec{v}_i = \vec{S}_i / \Delta t$ , but using a notation which emphasizes that the displacement  $\vec{S}_i = \Delta \vec{R}_i$  is the *change in position* that occurs during the time  $\Delta t$ . The Greek letter  $\Delta$  (delta) is used both to represent the idea that the quantity  $\Delta \vec{R}_i$  or  $\Delta t$ is small, and to emphasize that both of these quantities change as we change the strobe rate.

The limiting process in Figure (1) can be written in the form

$$\vec{\underline{v}}_{i} \equiv \frac{\text{Limit}}{\Delta t \to 0} \frac{\Delta \dot{R}_{i}}{\Delta t}$$
(3)

where the word "Limit" with  $\Delta t \rightarrow 0$  underneath, is to be read as "limit as  $\Delta t$  goes to zero". For example we would read Equation (3) as "*the instantaneous velocity*  $\vec{v}_i$  *at position* i *is the limit, as*  $\Delta t$  *goes to zero, of the ratio*  $\Delta \vec{R}_i / \Delta t$ ."

For two reasons, Equation (3) is not quite yet in standard calculus notation. One is that in calculus, only the limiting value, in this case, the instantaneous velocity, is considered to be important. Our strobe definition  $\vec{v}_i = \vec{\Delta R}_i / \Delta t$  is only a step in the limiting process. Therefore when we see the vector  $\vec{v}_i$ , we should assume that it is the limiting value, and no special symbol like the underline is used. For this reason we will drop the underline and write

$$\vec{v}_i = \frac{\text{Limit}}{\Delta t \to 0} \frac{\Delta \vec{R}_i}{\Delta t}$$
 (3a)

The second change deals with the fact that when  $\Delta t$  goes to zero we need an infinite number to time steps to get through our strobe photograph, and thus it is not possible to locate a position by counting time steps. Instead we measure the time t that has elapsed since the beginning of the photograph, and use that time to tell us where we are, as illustrated in Figure (4). Thus instead of using  $\vec{v}_i$  to represent the velocity at position i, we write  $\vec{v}(t)$  to represent the velocity at time t. Equation (3) now becomes

$$\vec{v}(t) = \frac{\text{Limit}}{\Delta t \to 0} \frac{\Delta \vec{R}(t)}{\Delta t}$$
 (3b)

where we also replaced  $\overrightarrow{\Delta R}_i$  by its value  $\overrightarrow{\Delta R}(t)$  at time t.

Although Equation (3b) is in more or less standard calculus notation, the notation is clumsy. It is a pain to keep writing the word Limit with a  $\Delta t \rightarrow 0$  underneath. To streamline the notation, we replace the Greek letter  $\Delta$  with the English letter d as follows

$$\frac{\operatorname{Limit}}{\Delta t \to 0} \frac{\overrightarrow{\Delta R}(t)}{\Delta t} \equiv \frac{d\vec{R}(t)}{dt}$$
(4)

(The symbol  $\equiv$  means *defined equal to*.) To a mathematician, the symbol  $d\vec{R}(t)/dt$  is just shorthand notation for the limiting process we have been describ-



Figure 4

Rather than counting individual images, we can locate a position by measuring the elapsed time t. In this figure, we have drawn the displacement vector  $\vec{R}(t)$  at time t = .3 sec. ing. But to a physicist, there is a different, more practical meaning. Think of dt as a short  $\Delta t$ , short enough so that the limiting process has essentially occurred, but not too short to see what is going on. In Figure (1), a value of dt less than .025 seconds is probably good enough.

If dt is small but finite, then we know exactly what the  $d\vec{R}(t)$  is. It is the small but finite displacement vector at the time t. It is our old strobe definition of velocity, with the added condition that dt is such a short time interval that the limiting process has occurred. From this point of view, which we will use throughout this text, dt is a real time interval, and  $d\vec{R}(t)$  a real vector which we can work with in a normal way. The only thing special about these quantities is that when we see the letter d instead of  $\Delta$ , we must remember that a limiting process is involved. In this notation, the calculus definition of velocity is

$$\vec{v}(t) = \frac{d\vec{R}(t)}{dt}$$
(5)

where R(t) and  $\vec{v}(t)$  are the particle's coordinate vector and velocity vector respectively as shown in Figure (5). Remember that this is just fancy shorthand notation for the limiting process we have been describing.



**Figure 5** *Instantaneous position and velocity at time t.* 

## ACCELERATION

In the analysis of strobe photographs, we defined both a velocity vector  $\vec{v}$  and an acceleration vector  $\vec{a}$ . The definition of  $\vec{a}$ , shown in Figure (2-12) reproduced here in Figure (6) was

$$\vec{a}_{i} \equiv \frac{\vec{v}_{i+1} - \vec{v}_{i}}{\Delta t}$$
(6)

In our graphical work we replaced  $\vec{v}_i$  by  $\vec{S}_i/\Delta t$  so that we could work directly with the displacement vectors  $\vec{S}_i$  and experimentally determine the behavior of the acceleration vector for several kinds of motion.

Let us now change this graphical definition of acceleration over to a calculus definition, using the ideas just applied to the velocity vector. First, assume that the ball reached position i at time t as shown in Figure (6). Then we can write

$$\vec{v}_i = \vec{v}(t)$$
  
 $\vec{v}_{i+1} = \vec{v}(t+\Delta t)$ 

to change the time dependence from a count of strobe flashes to the continuous variable t. Next, define the vector  $\Delta \vec{v}(t)$  by

$$\vec{\Delta v}(t) \equiv \vec{v}(t+\Delta t) - \vec{v}(t) \quad \left( = \vec{v}_{i+1} - \vec{v}_i \right)$$
(7)

We see that  $\overrightarrow{\Delta v}(t)$  is the change in the velocity vector as the time advances from t to  $t+\Delta t$ .

The strobe definition of  $\vec{a}_i$  can now be written

$$\vec{a}(t) \begin{pmatrix} strobe\\ definition \end{pmatrix} = \frac{\vec{v}(t + \Delta t) - \vec{v}(t)}{\Delta t} \equiv \frac{\vec{\Delta v}(t)}{\Delta t} \quad (8)$$

Now go through the limiting process, turning the strobe up, reducing  $\Delta t$  until the value of  $\vec{a}(t)$  settles down to its limiting value. We have

$$\vec{a}(t) \begin{pmatrix} calculus\\ definition \end{pmatrix} = \frac{\text{Limit}}{\Delta t \to 0} \left( \frac{\vec{v}(t + \Delta t) - \vec{v}(t)}{\Delta t} \right)$$

$$= \frac{\text{Limit}}{\Delta t \to 0} \left( \frac{\Delta \vec{v}(t)}{\Delta t} \right)$$
(9)

Finally use the shorthand notation d/dt for the limiting process:

$$\vec{a}(t) = \frac{d\vec{v}(t)}{dt}$$
(10)

Equation (10) does not make sense unless you remember that it is notation for all the ideas expressed above. Again, physicists think of dt as a short but finite time interval, and  $d\vec{v}(t)$  as the small but finite change in the velocity vector during the time interval dt. It's our strobe definition of acceleration with the added requirement that  $\Delta t$  is short enough that the limiting process has already occurred.



**Figure 6** *Experimental definition of the acceleration vector.* 

## Components

Even if you have studied calculus, you may not recall encountering formulas for the derivatives of vectors, like  $d\vec{R}(t)/dt$  and  $d\vec{v}(t)/dt$  which appear in Equations (5) and (10). To bring these equations into a more familiar form where you can apply standard calculus formulas, we will break the vector Equations (5) and (10) down into component equations.

In the chapter on vectors, we saw that any vector equation like

$$\vec{A} = \vec{B} + \vec{C} \tag{11}$$

is equivalent to the three component equations

$$A_x = B_x + C_x$$
  

$$A_y = B_y + C_y$$
  

$$A_z = B_z + C_z$$
(12)

The advantage of the component equations was that they are simply numerical equations and no graphical work or trigonometry is required. The limiting process in calculus does not affect the decomposition of a vector into components, thus Equation (5) for  $\vec{v}(t)$  and Equation (10) for  $\vec{a}(t)$  become

$$\vec{v}(t) = d\vec{R}(t)/dt$$
 (5)

$$v_x(t) = dR_x(t)/dt$$
 (5a)

$$v_{y}(t) = dR_{y}(t)/dt$$
 (5b)

$$v_z(t) = dR_z(t)/dt$$
 (5c)

and

$$\vec{a}(t) = d\vec{v}(t)/dt \tag{10}$$

$$a_{x}(t) = dv_{x}(t)/dt$$
(10a)

$$a_{y}(t) = dv_{y}(t)/dt$$
(10b)

$$a_z(t) = dv_z(t)/dt$$
 (10c)

Often we use the letter x for the x coordinate of the vector  $\vec{R}$  and we use y for  $R_y$  and z for  $R_z$ . With this notation, Equation (5) assumes the shorter and perhaps more familiar form

T

$$v_{x}(t) = dx(t)/dt \qquad y \qquad (5a')$$

$$v_{y}(t) = dy(t)/dt \qquad (5b')$$

$$v_{z}(t) = dz(t)/dt \qquad (5c')$$

At this point the notation has become deceptively short. You now have to remember that x(t) stands for the x coordinate of the particle at a time t.

We have finally boiled the notation down to the point where it would be familiar from any calculus course. If we restrict our attention to one dimensional motion along the x axis. Then all we have to concern ourselves with are the x component equations

$$v_{x}(t) = \frac{dx(t)}{dt}$$

$$a_{x}(t) = \frac{dv_{x}(t)}{dt}$$
(10a)

## Distance, Velocity and Acceleration versus Time Graphs

One of the ways to build an intuition for Equations (5a) and (10a) is through the use of graphs of position, velocity and acceleration versus time. Suppose, for example, we had a particle moving at constant speed in the x direction, the uniform motion that the principle of relativity tells us that we cannot detect. Graphs of distance x(t), velocity v(t) and acceleration a(t) for this motion are shown in Figure (7).



**Figure 7** *Motion with constant velocity.* 

If the particle is moving at constant speed

$$\mathbf{v}_{\mathbf{x}}(\mathbf{t}) = \mathbf{v}_0 \tag{11}$$

then the graph of velocity versus time is a straight horizontal line of height  $v_0$  as shown in Figure (7b).

If you travel away from home at constant speed, then your distance from home is proportional to the time you have traveled. If you start at t = 0, then at time t your distance from home is

$$\mathbf{x}(\mathbf{t}) = \mathbf{v}_0 \mathbf{t} \tag{12}$$

This is graphed as the straight line as shown in Figure (7a). The slope of this line, the tangent of the angle  $\theta$  is x/t, which from Equation (12) is v<sub>0</sub>.

When a particle moves at constant velocity, there is no change in the succeeding velocity vectors, thus the acceleration a(t) is zero for all time

$$\mathbf{a}(\mathbf{t}) = \mathbf{0} \tag{13}$$

as shown in Figure (7c).

In summary, we have seen that for this example of uniform motion in the x direction

$$x(t) = v_0 t$$
 (12)  
 $y(t) = v_0$  (11)

$$\mathbf{v}(\mathbf{t}) = \mathbf{v}_0 \tag{11}$$

$$\mathbf{a}(\mathbf{t}) = \mathbf{0} \tag{13}$$

Now let us see if these results agree with our calculus definitions (5a) and (10a). From Equation (5a) we get

$$\mathbf{v}(t) = \frac{\mathrm{d}\mathbf{x}(t)}{\mathrm{d}t} = \frac{\mathrm{d}}{\mathrm{d}t}(\mathbf{v}_0 t) \tag{14}$$

The  $v_0$  being constant comes outside and we have

$$\mathbf{v}(t) = \mathbf{v}_0 \frac{\mathrm{d}t}{\mathrm{d}t} = \mathbf{v}_0 \tag{15}$$

where we used dt/dt = 1. Our calculus result agrees with Equation (12).

From Equation (10), we get

$$a(t) = \frac{dv(t)}{dt} = \frac{d}{dt}(v_0) = 0$$
(16)

because the derivative of a constant is zero.



**Figure 8** *Motion with constant acceleration.* 

What we should begin to see from this example, is that if we have the formula for x(t) then it is easy to use calculus to figure out the particle's velocity and acceleration. Let us consider one more example. Suppose x(t) is given by the formula

$$\mathbf{x}(\mathbf{t}) = \mathbf{a}\mathbf{t} + \mathbf{b}\mathbf{t}^2 \tag{17}$$

where a and b are constants. Then the calculus formulas (5a) and (10a) give

$$v(t) = \frac{dx(t)}{dt} = a + b \frac{d}{dt} (t^2)$$
$$= a + 2bt$$
(18)

where we used  $d(t^2)/dt = 2t$ . Equation (10a) gives

$$a(t) = \frac{dv(t)}{dt} = 2b$$
(19)

The results in Equations (17), (18) and (19) are graphed in Figures (8a, b and c) The position vs time a straight line with a slope 2b, and the acceleration is a constant 2b. Figure (8) therefore represents an example of motion with constant acceleration.

## THE CONSTANT ACCELERATION FORMULAS

Unfortunately life is not as simple as one might think from the preceding example. If you have the formula for x(t), then you can calculate v(t) and a(t) very easily by differentiation. But usually you have to go the other way. From the physics you figure out what the acceleration is, then you have to work back to get v(t) and finally x(t). At best, this reverse process involves integration which is typically quite a bit harder than differentiation.

Let us work out an example where we know the acceleration and have to integrate to get the velocity and position. We will take the easiest non trivial case where the acceleration is constant. The result will be the constant acceleration formulas.

If we know a(t), the first step is to *solve* equation (10a) by turning it into an integral equation as follows

$$a(t) = \frac{dv(t)}{dt}$$
(10a)

First multiply both sides by dt. (Remember that physicists keep dt very small but finite, so that we can move it around.) We get

$$dv(t) = a(t)dt$$
(20)

Now integrate both sides of Equation (20) from time t=0 up to time t=T. (This is called a definite integral.) We get

$$\int_0^T d\mathbf{v}(t) = \int_0^T \mathbf{a}(t)dt \tag{21}$$

The integral on the left is simply v(t) evaluated between 0 and T.

$$\int_{0}^{T} dv(t) = v(t) \Big|_{0}^{T} = v(t) - v(0)$$
 (22)

On the right side of Equation (21), we set  $a(t) = a_0$  (for constant acceleration) to get

$$\int_{0}^{T} a(t)dt = \int_{0}^{T} a_{0}dt = a_{0}\int_{0}^{T} dt$$
$$= a_{0}t\Big|_{0}^{T} = a_{0}T - a_{0} \times 0 = a_{0}T$$
(23)

Using Equations (22) and (23) in (21) we get

$$v(T) - v(0) = a_0 T$$
 (24)

The next step is to recognize that Equation (24) applies to any time T, so that we can replace T by t to get

$$v(t) = v(0) + a_0 t$$
 (25)

To emphasize that v(0), the particle's speed at time t = 0, is not a variable, we will use the notation  $v(0) \equiv v_0$  and Equation (25) becomes

$$v(t) = v_0 + a_0 t$$
 (26)

(If the steps we have used to derive Equation (26) were familiar and comfortable, then your calculus background is in good shape and you should not have much of a problem with calculus in reading this text. If, on the other hand what we did was strange, if the notation was unfamiliar and the steps unpredictable, a review of calculus is indicated. What we have done in the derivation of Equation (26) is use the concept of a definite integral. We will use definite integrals throughout the course and now is the time to learn how to use them. You should also be sure that you can do simple differentiations like  $d/dt(at^2) = 2at$ .) To get the other constant acceleration formula, start with Equation (5a)

$$v(t) = \frac{dx(t)}{dt}$$
(5a)

and multiply through by dt to get

$$d\mathbf{x}(t) = \mathbf{v}(t)dt \tag{27}$$

Again integrate both sides from t = 0 to t = T to get

$$\int_0^T dx(t) = \int_0^T v(t)dt$$
(28)

We can immediately do the integral on the left hand side

$$\int_{0}^{T} dx(t) = x(t) \Big|_{0}^{T} = x(T) - x(0)$$
 (29)

At this point we cannot do the integral on the right side of Equation (28) until we know explicitly how v(t)depends on the variable t. If, however, the acceleration is constant, we can use Equation (26) for v(t) to get

$$\int_{0}^{T} v(t) dt = \int_{0}^{T} (v_0 + a_0 t) dt$$
  
=  $\int_{0}^{T} v_0 dt + \int_{0}^{T} a_0 t dt$  (30)  
=  $v_0 \int_{0}^{T} dt + a_0 \int_{0}^{T} t dt$ 

Knowing that

$$\int_{0}^{T} t \, dt = \frac{t^2}{2} \tag{31}$$

we get

$$\int_{0}^{T} v(t)dt = v_{0}t \Big|_{0}^{T} + a_{0}\frac{t^{2}}{2}\Big|_{0}^{T}$$

$$= v_{0}T + a_{0}\frac{T^{2}}{2}$$
(32)

Using Equations (29) and (32) in (28) gives

$$x(T) - x_0 = v_0 T + \frac{1}{2} a_0 T^2$$
 (33)

Since Equation (33) applies for any arbitrary time T, we can replace T by t to get

$$\mathbf{x}(t) = \mathbf{x}_0 + \mathbf{v}_0 t + \frac{1}{2} \mathbf{a}_0 t^2$$
(34)

where we have written  $x_0$  for x(0), the position of the particle at time t = 0.

## **Three Dimensions**

Equations (26) and (34) are the constant acceleration formulas for motion in one dimension, along the x axis. (We can, of course, choose the x axis to point any way we want.)

If we want to describe motion in three dimensions with constant acceleration, we repeat the steps leading to Equations (26) and (34), but starting with (5b) and (10b) for motion along the y axis, and (5c) and (10c) for motion along the z axis. The steps are essentially identical, and we end up with the six equations

$$x(t) = x_0 + v_x(0)t + \frac{1}{2}a_xt^2$$
 (35a)

$$y(t) = y_0 + v_y(0)t + \frac{1}{2}a_yt^2$$
 (35b)

$$z(t) = z_0 + v_z(0)t + \frac{1}{2}a_z t^2$$
 (35c)

$$v_x(t) = v_x(0)t + a_x t$$
 (36a)

$$v_{y}(t) = v_{y}(0)t + a_{y}t$$
 (36b)

$$\mathbf{v}_{\mathbf{z}}(\mathbf{t}) = \mathbf{v}_{\mathbf{z}}(0)\mathbf{t} + \mathbf{a}_{\mathbf{z}}\mathbf{t}$$
(36c)

where we have temporarily gone back to the notation x(0) for  $x_0$ ,  $v_x(0)$  for  $v_{x0}$ , etc., and  $a_x$ ,  $a_y$ , and  $a_z$  are the x, y, z components of the assumed *constant* acceleration.

In Chapter 3 we introduced a notation that allowed us to conveniently express a vector  $\vec{S}$  in terms of its components  $S_x$ ,  $S_y$  and  $S_z$ , by writing the components, separated by commas, inside a parenthesis as follows

$$\mathbf{S} \equiv (\mathbf{S}_{\mathbf{x}}, \mathbf{S}_{\mathbf{y}}, \mathbf{S}_{\mathbf{z}}) \tag{37}$$

Using this notation, we define the following vectors by their components

$$\vec{R}(t) \equiv \left(x(t), y(t), z(t)\right) \qquad \begin{array}{c} coordinate \\ vector \end{array} (38)$$

$$\vec{v}(t) \equiv \left(v_x(t), v_y(t), v_z(t)\right) \begin{array}{c} velocity \\ vector \end{array} (39)$$

$$\vec{a} \equiv (a_x, a_y, a_z)$$
   
 $\vec{a} \equiv (a_x, a_y, a_z)$ 
  
 $constant$ 
  
 $acceleration$ 
  
(40)

With this vector notation, the six constant acceleration formulas (35a, b, c) and (36 a, b, c) reduce to the two vector equations

$$\vec{x}(t) = \vec{x}(0) + \vec{v}(0)t + \frac{1}{2}\vec{a}t^2$$
 (35)

$$\vec{v}(t) = \vec{v}_0(0) + \vec{a}t$$
 (36)

or using the notation  $\vec{R}_0 = \vec{R}(0)$ ,  $\vec{v}_0 = \vec{v}(0)$ , we have

$$\vec{R}(t) = \vec{R}_0 + \vec{v}_0 t + \frac{1}{2}\vec{a}t^2$$
 (35')

$$\vec{v}(t) = \vec{v}_0 + \vec{a}t$$
 (36')

These are the set of vector equations that we tested in our studies in Chapter 3 of instantaneous velocity with constant acceleration.

We have gone through all the details of the derivation of Equations (35) and (36), because they represent one of the major successes of the use of calculus in the prediction of motion. Whenever a particle's acceleration  $\vec{a}$  is constant, and we know  $\vec{a}, \vec{R}_0$ , and  $\vec{v}_0$ , we can use these equations to predict the particle's position  $\vec{R}(t)$ and velocity  $\vec{v}(t)$  at any time t in the future.

## **PROJECTILE MOTION** WITH AIR RESISTANCE

In our experimental study of projectile motion, we saw that when we used a styrofoam projectile, air resistance affected the acceleration of the projectile. From the point of view that we are riding on the ball, we would feel a wind in our face, blowing in a direction  $-\vec{v}$ , opposite to the velocity  $\vec{v}$  of the projectile. The effect of this wind was to blow the acceleration vector back as shown in Figure (3-28), reproduced here as Figure (9).

We saw that the experimental vector  $\vec{a}_3$  was the acceleration  $\vec{g}$  we would have in the absence of air resistance, plus a correction  $\vec{a}_{air}$  which pointed in the direction of the wind, in the  $-\vec{v}$  direction as shown.

The magnitude  $|\vec{a}_{air}|$  cannot accurately be determined from the strobe photograph. About all we can tell is that  $|\vec{a}_{air}|$  is zero if the ball is at rest, and increases as the speed  $|\vec{v}|$  of the ball increases. The simplest guess is that  $|\vec{a}_{air}|$  is proportional to  $|\vec{v}|$  and we have the formula

simple

 $\vec{a} \cdot = -K\vec{v}$ 

$$\vec{a}_{air} = -K\vec{v} \qquad simple \\ guess \qquad (41)$$

$$\vec{a}_{3} \qquad \vec{v}_{vind''} \\ \vec{a}_{air} \qquad \vec{v}_{vind''} \\ \vec{a}_{air} = -K\vec{v}$$
Figure 9



Our strobe photograph does not eliminate the possibility that  $|\vec{a}_{air}|$  is more complicated, something like

$$\vec{a}_{air} = K_2 |\vec{v}|^2 \tag{42}$$

or perhaps some combination like

$$|\vec{a}_{air}| = K_1 |\vec{v}| + K_2 |\vec{v}|^2$$
 (43)

It turns out that the motion of a sphere through a liquid (in our case a Styrofoam ball through air) has been studied extensively by both physicists and engineers. For slow speeds the motion is like Equation (41) but as the speed increases it looks more like Equation (43) and soon becomes even more complicated. The only simple fact is that  $|\vec{a}_{air}|$  always points in the direction  $-\vec{v}$ , in the direction of the wind in our face (until vortex shedding occurs).

As an exercise to test the ability of calculus to predict motion, let us assume that our simple guess  $\vec{a}_{air} = -K\vec{v}$ is good enough. We would then like to solve the calculus Equations (5) and (10) for the case where the acceleration is not constant, but is given by the formula

$$\vec{a} = \vec{g} - K\vec{v} \tag{44}$$

where  $\vec{g}$  is the constant acceleration due to gravity,  $\vec{v}$  is the instantaneous velocity of the particle, and K is what we will call the air resistance constant. Equation (44) is pictured in Figure (9).



Figure 10

Our first step is to introduce a coordinate system as shown in Figure (10), and break the motion up into x and y components. Since the acceleration  $\vec{g}$  due to gravity points down, we have  $g_x = 0$  and the vector Equation (44) can be written as the two component equations

$$a_{x} = -K v_{x} \qquad \left(g_{x} = 0\right) \qquad (44a)$$

$$a_y = g - K v_y$$
  $(g = -980 \text{ cm/sec}^2)$  (44b)

The calculus Equations (10a, b) that we have to solve become

$$a_{x} = \frac{dv_{x}}{dt} = -Kv_{x}$$
(45)

$$a_y = \frac{dv_y}{dt} = g - Kv_y$$
(46)

Let us focus on the simpler of the two equations, Equation (45) for the horizontal velocity of the projectile. We want to solve the equation

$$\frac{\mathrm{d}\mathbf{v}_{\mathbf{x}}(t)}{\mathrm{d}t} + \mathbf{K}\mathbf{v}_{\mathbf{x}}(t) = 0 \tag{45'}$$

Suppose we try to solve Equation (45) using the same steps we used to predict  $v_x$  for constant acceleration (Equations 20 through 26). Multiplying through by dt gives

$$dv_{x}(t) = -Kv_{x}(t)dt$$
(47)

Integrating from t = 0 to t = T gives

$$\int_{0}^{T} dv_{x}(t) = -\int_{0}^{T} K v_{x}(t) dt$$
 (48)

We can do the integral on the left, and remove the K from the integral on the right giving

$$v_x(t)\Big|_0^T = v_x(T) - v_x(0) = -K \int_0^T v_x(t) dt$$
 (49)

Now we are in trouble, because we have to integrate  $v_x(t)$  in order to find  $v_x(t)$ . We can't do the integral until we know the answer, and we have to do the integral to get the answer. It boils down to the fact that the techniques we used to solve the calculus equations for constant acceleration do not work now. As soon as the acceleration is not constant, we have a much more difficult problem.

## DIFFERENTIAL EQUATIONS

Equation (45) is an example of what is called a *differential equation*. (An equation with derivatives in it.) Only in very special cases, as in our example of constant acceleration, can these equations be solved in a straightforward manner by integration. In slightly more complicated cases, these equations can be solved by certain standard tricks that one learns in an advanced calculus course on differential equations. We will use one of these tricks to solve Equation (45).

In general, however, differential equations cannot be solved without numerical methods that are now handled by digital computers. If, for example we assumed that the air resistance was proportional to  $v^2$  as in Equation (42), then Equation (45) for the x component of velocity would be replaced by

$$\frac{dv_{x}(t)}{dt} + K_{2}v_{x}(t)^{2} = 0$$
(45a)

Equation (45a) is what is called a *non linear* differential equation, the word non linear coming from the appearance of the square of the unknown variable  $v_x(t)$ .

At the current time, there is no general way to solve non linear differential equations except by computer. Non linear differential equations have marvelously complicated features like *chaotic behavior* that have been discussed extensively in the popular press in the last few years. It is currently a hot research topic.

The point of this discussion is that when we use calculus to predict motion, a very slight increase in the complexity of the problem can lead to enormous increases in the difficulty in solving the problem. When the projectile's acceleration was constant, we could easily solve the calculus equations to get the constant acceleration formulas. If the air resistance has the simple form  $\vec{a}_{air} = -K\vec{v}$ , then we have to solve a differential equation, but we can still get an answer, a formula that predicts the motion of the particle. If we go up one step in complexity, if  $|\vec{a}_{air}|$  is proportional to the square of the speed, then we have a non linear differential equation that we cannot solve without numerical or approximation techniques.

Calculus gives marvelous results when we can solve the problem. We get formulas describing the motion at all future times. But we are extremely limited in the kind of problems that can be solved. Simple physical modifications of a problem can turn an easy problem into an unsolvable one.

Before inventing calculus, Isaac Newton invented a simple step-by-step method that we will discuss in the next chapter. Newton's step-by-step method has the great advantage that slight complications in the physical setup lead to only slightly more work in obtaining a solution. We will see that it is almost no harder to predict projectile motion with air resistance, even with  $v^2$  terms, than it is to predict projectile motion with air will allow us to handle problems in this course, realistic problems, that do not have a calculus solution.

There are two disadvantages to the step-by-step method, however. One, is that you get a numerical answer, like an explicit orbit, rather than a general result. In contrast, the constant acceleration formulas describe all possible trajectories for motion with constant acceleration.

The second problem is that in the step-by-step method, a simple calculation is repeated many times, perhaps thousands or millions of times to obtain an accurate answer. Before digital computers, lifetimes were spent doing this kind of calculation by hand to predict the motion of the moon. But modern digital computers have changed all that. In minutes, the digital computer running your word processor can do what used to be months of work.

## **Solving the Differential Equation**

We have essentially finished what we wanted to say about applying calculus to the problem of projectile motion with air resistance. The gist is that adding air resistance turns a simple problem into a hard one. Even for the simplest form of air resistance,  $\vec{a}_{air} = -K\vec{v}$ , we end up with the differential equation

$$\frac{\mathrm{d}\mathbf{v}_{\mathbf{x}}(t)}{\mathrm{d}t} + \mathbf{K}\mathbf{v}_{\mathbf{x}}(t) = 0 \tag{45}$$

which cannot be solved directly with integration.

Later in the course we will encounter several other differential equations, one having the same form as Equation (45). When we meet these equations, we will show you how to solve them.

At this time, we do not really need the solution to Equation (45). This equation does not represent a basic physics problem because our formula for air resistance is an approximation of limited validity. We include a solution for those who are interested, who want to see the problem completed now. Those for whom calculus is new or rusty may wish to skip to the next chapter.

The reason that differential equations are hard to solve is that the solutions are curves or functions rather than numbers. For example, the solution to the ordinary Equation  $x^2 = 4$  is the pair of numbers x = +2 and x =- 2. But Equation (45) has the decaying exponential curve shown in Figure (11) for a solution. What this curve tells us is that the v<sub>x</sub> or the horizontal motion, dies out in time and the projectile will eventually have only y motion. After enough time the ball will be falling straight down.

One of the standard techniques for solving differential equations is to guess the answer and then plug your guess into the equation to see if you are right. When you take a course in solving differential equations, you learn how to make educated guesses. If you had been through such a course, you would guess that Equation (45) should have an exponentially decaying solution, and try a solution of the form

$$v_x(t) = v_{x0} e^{-\alpha t} \qquad (guess) \tag{50}$$

where  $\alpha$  and  $v_{x0}$  are constants whose values we wish to find.



**Figure 11** Air resistance causes the horizontal component of the velocity to decay exponentially.

Differentiating Equation (50) gives

$$\frac{\mathrm{d}\mathbf{v}_{\mathbf{x}}(t)}{\mathrm{d}t} = -\mathbf{v}_{\mathbf{x}0}\alpha \mathrm{e}^{-\alpha t}$$
(51)

where we used the fact that

$$\frac{\mathrm{d}\mathrm{e}^{-\alpha \mathrm{t}}}{\mathrm{d}\mathrm{t}} = -\alpha \mathrm{e}^{-\alpha \mathrm{t}} \tag{52}$$

Substituting Equations (50) and (51) into Equation (45) gives

$$-v_{x0}\alpha e^{-\alpha t} + K v_{x0} e^{-\alpha t} = 0$$
 (53)

First note that the exponential function  $e^{-\alpha t}$  cancelled out. This indicates that we have guessed the correct function.

Next note that  $v_{x0}$  cancels. This means that any value of  $v_{x0}$  in Equation (50) is a possible solution. The particular value we want will be determined by the experimental situation.

What we have left is

$$\alpha = K \tag{54}$$

Thus the differential Equation (45) has the exponentially decaying solution

$$v_x(t) = v_{x0} e^{-Kt}$$
 (55)

where the decay rate is the air resistance constant K.

For those of you who have actually had a course in solving differential equations, see if you can solve for the vertical motion of the projectile. The differential equation you have to solve is

$$\frac{\mathrm{d}\mathbf{v}_{\mathbf{y}}(\mathbf{t})}{\mathrm{d}\mathbf{t}} = \mathbf{g} - \mathbf{K}\mathbf{v}_{\mathbf{y}}(\mathbf{t}) \tag{46}$$

The answer for long times turns out to be simple – the projectile ends up coasting at a constant terminal velocity. See if you can get that result.

The answer is  $v_y = (g/K)(1 - e^{-Kt})$  if  $v_y = 0$  at t = 0.

# Appendix A

## SOLVING PROJECTILE MOTION PROBLEMS

In high school physics texts and most college level introductory physics texts, there is considerable emphasis on solving projectile motion problems. A good reason for this is that these problems provide practice in problem solving techniques such as drawing clear sketches, developing an orderly approach, and checking units. Not such a good reason is that, in texts that rely solely on algebra and calculus, the only thing they can solve in the early stages are projectile motion or circular motion problems.

The disadvantage of over emphasizing projectile motion problems is that students begin to use the projectile motion formulas as a general way of predicting motion, using the formulas in circumstances where they do not apply. The important point to remember is that the formulas  $\vec{v} = \vec{v}_i + \vec{a}t$  and  $\vec{x} = \vec{v}_i t + 1/2\vec{a}t^2$  are very limited in scope. They apply only when the acceleration  $\vec{a}$  is constant, a not very likely circumstance in the real world. The acceleration  $\vec{a}$  is not constant for circular motion, projectile motion with air resistance, satellite motion, the motion of electrons in a magnetic field, and most interesting physics problems.

From the point of view that solving projectile motion problems is basically for practice in problem solving techniques, we will show you an orderly way of handling these problems. The approach which we will illustrate using several examples should allow you with practice to handle any constant acceleration problems test makers throw at you. In these examples, we are demonstrating not only how the problem is solved but also how you should go about doing it.

(Note -- in this appendix, all velocity vectors are instantaneous velocities, thus we will not bother underlining them.)

#### Example A1

A boy throws a ball straight up into the air and catches it (at the same height from which he threw it) 2 sec later. How high did the ball go?

**Solution**: To solve all projectile problems, we use the equations

$$\vec{S} = \vec{v}_i t + \frac{1}{2} \vec{a} t^2$$
$$\vec{v}_f = \vec{v}_i + \vec{a} t$$

However these are vector equations. Using a coordinate system in which the y axis is in the vertical direction and the x axis is in the horizontal direction, we get the following equations.

## Vertical motion:

$$S_y = v_{iy}t + \frac{1}{2}a_yt^2$$
$$v_{fy} = v_{iy} + a_yt$$

Horizontal motion:

$$S_x = v_{ix}t + \frac{1}{2}a_xt^2$$
$$v_{fx} = v_{ix} + a_xt$$

Now projectiles near the surface of the earth accelerate downward at a rate of nearly 980 cm/sec<sup>2</sup>. This value varies slightly at different points on the surface of the earth, but is always quite close to 980 cm/sec<sup>2</sup>. This acceleration due to gravity is usually designated g; since it is directed downward in the *minus* y direction, we have

$$a_y = -g = -980 \text{ cm/sec}^2 (-32 \text{ ft/sec}^2)$$
  
 $a_x = 0$ 

As a result, we get the equations

Vertical motion:

(a) 
$$S_y = v_{iy}t - \frac{1}{2}gt^2$$
 (A1a)

(b) 
$$v_{fy} = v_{iy} - gt$$
 (A1b)

Horizontal motion:

(c) 
$$S_x = v_{ix}t$$
 (A1c)

(d) 
$$v_{fx} = v_{ix}$$
 (A1d)

Horizontal motion and vertical motion are entirely independent of each other. We see, for example, from Equation(A1d), that the horizontal speed of a projectile does not change; but this has already been obvious from the strobe photographs.

Now let us apply Equation (A1) to the situation where the boy throws the ball straight up and catches it 2 sec later. Since there is no horizontal motion, we only need equations (A1a, b).

One good technique for solving projectile problems is to work up to and back from the top of the trajectory. The reason is that at the top of the trajectory, Equations (A1a, b) are very easily applied.

In our problem, the ball spent half its time going up and half its time falling; thus, the fall took 1 sec. The distance that it fell is

$$S_y = v_{iy}t - \frac{1}{2}gt^2$$

where t is 1 sec, and since we are starting at the top of the trajectory. We get

$$S_y = -\frac{1}{2}gt^2 = -\frac{1}{2} \times 32 \text{ ft/sec}^2 \times 1 \text{ sec}^2$$
$$S_y = -16 \text{ ft}$$

The minus sign indicated the ball fell 16 ft below the top of the trajectory.

#### Example A2

A ball is thrown directly upward at a speed of 48 ft/sec. How high does it go?

**Solution:** First, find the time it takes to reach to top of its trajectory. We have

$$v_{iy} = 48 \text{ ft/sec}$$

$$v_{fy} = 0$$
 at the top of the trajectory

From Equation (4-A1b) we have

$$v_{fy} = v_{iy} - gt$$

or

$$t = \frac{v_{iy}}{g} = \frac{48 \text{ ft/sec}}{32 \text{ ft/sec}^2} = 1.5 \text{ sec}$$

Now we can use Equation A1a to calculate how high the ball goes.

$$S_y = v_{iy}t - \frac{1}{2}gt^2$$

We have  $v_{iy} = 48$  ft/sec, t = 1.5 sec to reach the top; thus S<sub>v</sub>, the distance to the top, is

$$S_y = 48 \frac{\text{ft}}{\text{sec}} \times 1.5 \text{ sec} - \frac{1}{2} \times 32 \frac{\text{ft}}{\text{sec}^2} \times (1.5 \text{ sec})^2$$
  
 $S_y = 72 \text{ ft} - 36 \text{ ft} = 36 \text{ ft}$ 

#### Example A3

An outfielder throws a ball at a speed of 96 ft/sec at an angle of 30° above the horizontal. How far away from the outfielder does the ball strike the ground?

**Solution:** When solving problems, the first step is to draw a neat diagram of the situation, as in Figure (A1). The first calculation is to find the x and y components of  $\vec{v}_i$ . From our diagram we see that

 $v_{ix} = v_i \cos \theta$ = 96 ft/sec × 0.864 = 83 ft/sec  $v_{iy} = v_i \sin \theta$ = 96 ft/sec × 0.50 = 48 ft/sec

where  $\cos 30^{\circ} = 0.864$  and  $\sin 30^{\circ} = 0.50$ .

Now we are in a position to separate the problem into two parts – vertical motion and horizontal motion. These may be treated as two independent problems.



#### Figure A1

Sketch of the problem. On the sketch, label the symbols used, show what is given, and state what you are to find. It is generally better to work the problem in terms of letters, substituting numbers only at the end, or at convenient breaks in the problem. *Vertical motion.* A ball is thrown straight up at a speed  $v_{iy} = 48$  ft/sec; how long a time t does it take to come back to the ground?

*Horizontal motion.* A ball travels horizontally at a speed  $v_{ix} = 83$  ft/sec. If it travels for a time t (result of vertical motion problem) how far does it travel?

We see that the vertical motion problem is exactly the one we solved in Example A-2,  $v_{iy} = 48$  ft/sec in both cases. Thus, using the same solution, we find that the ball takes 1.5 sec to go up and another 1.5 sec to come down, for a total time of

$$t = 3 \text{ sec}$$

Now solve the horizontal motion from

$$S_x = v_{ix}t$$

We get

$$S_x = 83 \text{ ft/sec} \times 3 \text{ sec} = 249 \text{ ft}$$

which is the answer.

#### **Checking Units**

It is easy to make a mistake when working a problem. One of the best ways to avoid mistakes is to write out the dimensions of each number used in the calculation; if the answer has the wrong dimension, you will know there is a mistake somewhere. For example, in the preliminary edition of this text the following formula accidentally appeared.

$$\vec{S} = \vec{v}_i + \frac{1}{2}\vec{a}t^2$$

Putting in the dimensions, we find

$$\vec{S} ft = \vec{v}_i \frac{ft}{sec} + \frac{1}{2}\vec{a}\frac{ft}{sec^2} \times (t sec)^2$$

or

$$\left(\vec{S}\right)ft = \left(\vec{v}_i\right)\frac{ft}{\sec} + \left(\frac{1}{2}\vec{a}t^2\right)ft$$

Clearly the  $(\vec{v}_i)$  ft/sec has the wrong dimensions, since we cannot add ft/sec to ft. Thus, through a check of the dimensions we would immediately spot an error in this formula, even if we had no idea what the formula is about. To correct this formula, the  $\vec{v}_i$  must be multiplied by t sec so that the result is  $(\vec{v}_i)$  ft/sec×t sec equals  $(\vec{v}_i t)$  ft.

As another instance, in the solution of Example (A3) we had

$$t = \frac{v_{iy}}{g}$$

At this point you might begin to worry that you have made a mistake; your doubts will be dispelled, however, once dimensions are inserted

t sec = 
$$\frac{v_{iy} \text{ ft/sec}}{\text{g ft/sec}^2} = \frac{v_{iy}}{\text{g}} \text{ sec}$$

#### **Exercise A1**

A 22-caliber rifle with a muzzle velocity of 600 ft/sec is fired straight up. How high does the bullet go? How long before it hits the ground?

## Exercise A2

(The rifle of Exercise A1 is fired at an angle of 45°. How far does the bullet travel? (Give answer in ft and in mi.)

#### **Exercise A3**

A right fielder is 200 ft from home plate. Just at the time he throws the ball into home plate, a runner leaves third base and takes 3.5 sec to reach home plate. If the maximum height reached by the ball is 64 ft, did the runner make it to home plate in time? (Problem from J. Orear, *Fundamental Physics*, Wiley, New York, 1961.)

#### **Exercise A4**

A steel ball is bouncing up and down on a steel plate with a period of oscillation of 1 sec. How high does it bounce? (Problem from J. Orear, *Fundamental Physics*, Wiley, New York, 1961.)

#### **Exercise A5**

A small rocket motor is capable of providing an acceleration of 0.01 g to a space capsule. If the capsule starts from a far-out space station and the rocket motor runs continuously, how far away is the capsule at the end of 1 year? What is the capsule's speed relative to the space station at the end of the year?

#### **Exercise A6**

A car traveling at 60 mi/hr strikes a tree. Inside the car the driver travels 1 ft from the time the car struck the tree until he is at rest. What is the deceleration of the driver if his deceleration is constant? Give the answer in ft/sec<sup>2</sup> and in g's.

## **Exercise A7**

During volcanic eruptions, chunks of solid rock can be blasted out of the volcano. These projectiles are called volcanic blocks. Figure (2) shows a cross-section of Mt. Fuji, in Japan.

At what initial speed  $v_0$  would a block have to be ejected, at  $45^\circ,$  in order to fall at the foot of the volcano as shown.

What is the time of flight?

(Problem from Halliday and Resnick.)

Hint Use the vector equation

$$\vec{S} = \vec{v}_1 t + \frac{1}{2} \vec{a} t^2$$
  
which is illustrated in Figure 3-34 reproduced to the  
right. In this problem  $\vec{S}$  is the total displacement of the  
rock, from the time it left the volcano until it hit the  
ground. Separate the vector equation into x and y  
components.  
 $\vec{A} = \frac{45^\circ}{45^\circ}$   
 $\vec{A} = \frac{45^\circ}{1000}$   
 $\vec{A} = \frac{1000}{1000}$   
 $\vec{A} = \frac{10000}{1000}$   
 $\vec{A} = \frac{1000}{1000}$   
 $\vec{A} = \frac{100$ 

₹Į,t

#### Figure A2

The farthest out blocks are the ones ejected at the greatest speed  $v_0$  at an angle of 45°. By noting that the most distant blocks are 9.4 km away, you can thus determine the maximum speed at which the blocks were ejected.

## Chapter 5 Computer Prediction of Motion

## **STEP-BY-STEP CALCULATIONS**

In the last chapter we saw that for the special case of constant acceleration, calculus allowed us to obtain a rather remarkable set of formulas that predicted the object's motion for all future times (as long as the acceleration remained constant). We ran into trouble, however, when the situation got a bit more complicated. Add a little air resistance and the analysis using calculus became considerably more difficult. Only for the very simplest form of air resistance are we able to use calculus at all.

On the other hand, adding a little air resistance had only a little effect on the actual projectile motion. Without air resistance the projectile's acceleration vectors pointed straight down and were all the same length, as seen in Figure (3-27). Include some air resistance using the Styrofoam projectile, and the acceleration vectors tilted slightly as if blown back by the wind one would feel riding along with the ball, as seen in Figure (3-31). Since projectile motion with air resistance is almost the same as that without, one would like a method of predicting motion that is almost the same for the two cases, a method that becomes only a little harder if the physical problem becomes only a little more complex.

The clue for developing such a method is to note that in our analysis of strobe photographs, we have been breaking the motion into short time intervals of length  $\Delta t$ . During each of these time intervals, not much happens. In particular, the Styrofoam projectile's acceleration vector did not change much. Only over the span of several intervals was there a significant change in the acceleration vector. This suggests that we could predict the motion by assuming that the Styrofoam ball's acceleration vector was essentially constant during each time interval, and at the end of each time step correct the acceleration vector in order to predict the motion for the next time step. In this way, by a series of short calculations, we can predict the motion over a long time period. This is a rough outline of the step-by-step method of predicting motion that was originally developed by Isaac Newton and that we will discuss in this chapter.

The problem with the step-by-step prediction of motion is that it quickly gets boring. You are continually repeating the same calculation with only a small change in the acceleration vector. Worse yet, to get very accurate results you should take very many, very small, time steps. Each calculation is almost identical to the previous one, and the process becomes tedious. If these calculations are done by hand, one needs an enormous incentive in order to obtain meaningful results.

## COMPUTER CALCULATIONS

Because of the tedium involved, step-by-step calculations were used only in desperate circumstances until the invention of the digital computer in the middle of the twentieth century. The digital computer is most effective and easiest to use when we have a repetitive calculation involving many, very similar steps. It is the ideal device for handling the step-by-step calculations described above. With a digital computer we can use very small time steps to get very accurate results, doing thousands or millions of steps to predict far into the future. We can cover the same range of prediction as the calculus-derived formulas, but not encounter significant difficulties when there is a slight change in the problem, such as the addition of air resistance.

To illustrate how to use the computer to handle a repetitive problem, we will begin with the calculation and plotting of the points on a circle. We will then go back to our graphical analysis of strobe photographs and see how that analysis can be turned into a series of steps for a computer prediction of motion.



Points on a circle.

## Calculating and Plotting a Circle

Figure (1) shows 100 points on the circumference of a circle of radius r. To make this example somewhat similar to the analysis of strobe photographs, we will choose a circle of radius r = 35 cm, centered at x = 50, y = 50, so that the entire circle will fit in the region x = 0 to 100, y = 0 to 100, as shown. The i th point around the circle has x and y coordinates given by

$$\begin{aligned} \mathbf{x}_i &= \mathbf{r} \cos \theta_i \\ \mathbf{y}_i &= \mathbf{r} \sin \theta_i \end{aligned} \tag{1}$$

where  $\theta_i$ , the angle to the i th point, is given by

$$\theta_{i} = \frac{360}{100} * i \text{ degrees} = \frac{2\pi}{100} * i \text{ radians}$$

(We know that it is easier to draw a circle using a compass than it is to calculate and plot all these individual points. But if we want something more complicated than a circle, like an ellipse or Lissajous figure, we cannot use a compass. Then we have to calculate and plot individual points as we are doing.)

If we wrote out the individual steps required to calculate and plot these 100 points, the result might look like the following:

$$\theta_0 = (2\pi/100) * 0 = 0 \text{ radians}$$
  

$$x_0 = 50 + r \cos(\theta_0) = 50 + 35 \cos(0)$$
  

$$= 50 + 35 * 1 = 85$$
  

$$y_0 = 50 + r \sin(\theta_0) = 50 + 35 \sin(0)$$
  

$$= 50 + 35 * 0 = 50$$
  
Plot a point at (x = 85, y = 50)

$$i = 1$$
  

$$\theta_1 = (2\pi/100) * 1 = .0628 \text{ radians}$$
  

$$x_1 = 50 + r \cos(\theta_1) = 50 + 35 \cos(.0628)$$
  

$$= 50 + 35 * .9980 = 84.93$$
  

$$y_1 = 50 + r \sin(\theta_1) = 50 + 35 \sin(.0628)$$
  

$$= 50 + 35 * .0628 = 52.20$$

Plot a point at (x = 84.93, y = 52.20)

•••

i = 50

$$\theta_{50} = (2\pi/100) * 50 = \pi \text{ radians}$$
  

$$x_{50} = 50 + r \cos(\theta_{50}) = 50 + 35 \cos(\pi)$$
  

$$= 50 + 35 * (-1) = 15$$
  

$$y_{50} = 50 + r \sin(\theta_{50}) = 50 + 35 \sin(\pi)$$
  

$$= 50 + 35 * 0 = 50$$

Plot a point at (x = 15, y = 50)

•••

In the above, not only will it be tedious doing the calculations, it is even tedious writing down the steps. That is why we only showed three of the required 100 steps.

The first improvement is to find a more efficient way of writing down the steps for calculating and plotting these points. Instead of spelling out all of the details of each step, we would like to write out a short set of instructions, which, if followed carefully, will give us all the steps indicated above. Such instructions might look as follows:

- 1) Let r = 35
- 2) Start with i = 0
- 3) Let  $\theta_i = (2\pi/100) * i$
- 4) Let  $x_i = 50 + r \cos \theta_i$
- 5) Let  $y_i = 50 + r \sin \theta_i$
- 6) Plot a point at  $x_i, y_i$
- 7) Increase i by 1
- 8) If i is less than 100, then go back to step 3 and continue in sequence
- 9) If you got here, i = 100 and you are done

#### Figure 2

A program for calculating the points around a circle.

#### Exercise 1

Follow through the instructions in Figure (2) and see that you are actually creating the individual steps shown earlier.

## **PROGRAM FOR CALCULATION**

The set of instructions shown in Figure (2) could be called a plan or *program* for doing the calculation. A similar set of instructions typed into a computer is called a *computer program*. Our instructions in Figure (2) would not be of much use to a person who spoke only German. But if we translated the instructions into German, then the German speaking person could follow them. Similarly, this particular set of instructions is not of much use to a computer, but if we translate them into a language the computer "understands", the computer can follow the instructions.

The computer language we will use in this course is called BASIC, a language developed at Dartmouth College for use in instruction. The philosophy in the design of BASIC is that it be as much as possible like an ordinary spoken language so that students can concentrate on their calculations rather than worry about details of operating the computer. Like human languages, the computer language has evolved over time, becoming easier to use and clearer in meaning. The version of BASIC we will use is called **True BASIC**, a modern version of BASIC written by the original developers of the language.

The way we will begin teaching you the language BASIC is to translate the set of instructions in Figure (2) into BASIC. We will do this in several steps, introducing a few new ideas at a time, just as you learn a few rules of grammar at a time when you are learning a foreign language. We will know that we have arrived at the actual language BASIC when the computer can successfully run the program. It is not unlike testing your knowledge of a foreign language by going out in the street and seeing if the people in that country understand you.

## The DO LOOP

In a sense, the set of instruction in Figure (2) is already in the form of sloppy BASIC, or you might say pidgin BASIC. We only have to clean up a few grammatical rules and it will work well. The first problem we will address is the statement in instruction #8.

8) If i is less than 100, then go back to step 3 and continue in sequence

There are two problems with this instruction. One is that it is long and wordy. Computer languages are usually designed with shorter, crisper instructions. The second problem is that the instruction relies on numbering instructions, as when we say "go back to step 3". There is no problem with numbering instructions in very short programs, but clarity suffers in long programs. The name "step 3" is not a particularly descriptive name; it does not tell us why we should go back there and not somewhere else. It is much better to state that we have a cyclic calculation, and that we should go back to the beginning of this particular cycle.

The grammatical construction we will use, one of the variations of the so-called "DOLOOP", has the following structure. We mark the beginning of the cyclic process with the word "DO", and end it with the command "LOOP UNTIL...". Applied to our instructions in Figure (2), the DO LOOP would look as follows:

LET r = 35  
LET i = 0  
DO  
LET 
$$\theta_i = (2\pi/100) * i$$
  
LET  $x_i = 50 + r \cos \theta_i$   
LET  $y_i = 50 + r \sin \theta_i$   
Plot a point at  $x_i, y_i$   
Increase i by 1  
LOOP UNTIL i = 100

All done

## Figure 3

Introducing the DO LOOP.

In the instructions in Figure (3), we begin by establishing that r = 35 and that i will start with the value 0. Then we mark the beginning of the cyclic calculation with the command **DO**, and end it with the command "LOOP UNTIL i = 100". The idea is that we keep repeating all the stuff between the "DO" line and the "LOOP..." line until our value of i has been incriminated up to the value i = 100. When i reaches 100, then the loop command is ignored and we have finished both the loop and the calculation.

## The LET Statement

Another major grammatical rule is needed before Figure (3) becomes a BASIC program that can be read by the computer. That involves a deeper understanding of the **LET** statement that appears in many of the instructions.

One example of a LET statement is the following

$$LET i = i + 1 \tag{2}$$

At first sight, statement (2) looks a bit peculiar. If we think of it as an equation, then we would cancel the i's and be left with

LET = 1

which is clearly nonsense. Thus the LET statement is not really an equation, and we have to find out what it is.

The LET statement combines the computer's ability to do calculations and to store numbers in memory. To understand the memory, think of the mail boxes at the post office. Above each box there is a name like "*Jones*", and Jones' mail goes inside the box. In the computer, each memory cell has a name like "i", and a number goes inside the cell. Unlike a mail box, which can hold several letters, a computer memory cell can store only one number at a time.

The rule for carrying out a LET statement like

LET i = i + 1

5-5

is to first evaluate the right hand side and store the results in the memory cell mentioned on the left side. In this example the computer evaluates i + 1 by first looking in cell "i" to see what number is stored there. It then adds 1 to that stored value to get the value (i + 1). To finish the command, it looks for a cell labeled "i", removes the number stored there and replaces it with the value just calculated. The net result of all this is that the numerical value stored in cell i is increased by 1.

There is a good mnemonic that helps you remember how a LET statement works. In the command LET i = i + 1, the computer takes the old value of i, adds one to get the new value, and stores that in cell i. If we write the LET statement as

LET 
$$i_{new} = i_{old} + 1$$

then it is clear what the computer is doing, and we are not tempted to cancel the i's. In this text we will often use the subscripts "old" and "new" to remind us what the computer is to do. When we actually type in the commands, we will omit the subscripts "old" and "new", because the computer does that automatically when performing a LET command.

With this understanding of the **LET** statement, our program for calculating the points on a circle becomes

LET r = 35  
LET i = 0  
DO  
LET 
$$\theta_i = (2\pi/100) * i$$
  
LET  $x_i = 50 + r \cos \theta_i$   
LET  $y_i = 50 + r \sin \theta_i$   
Plot a point at  $x_i, y_i$   
LET  $i_{new} = i_{old} + 1$   
LOOP UNTIL i = 100

All done

## Figure 4

Handeling the LET statement.

In Figure (4), we begin our repetitive **DO LOOP** by calculating a new value of the angle  $\theta$ . This new value is stored in the memory cell labeled  $\theta$ , and later used to calculate new values of  $x = r \cos \theta$  and  $y = r \sin \theta$ . Since we are using the updated values of

 $\theta$ , we can drop the subscripts i on the variables  $\theta_i$ ,

 $x_i, y_i$ . After we plot the point at the new coordinate (x, y), we calculate the next value of i with the command LET i = i + 1, and then go back for the next calculation.

To get a working BASIC program, there are a few other small changes that are easily seen if we compare our program in Figure (4) with the working BASIC program in Figure (5). Let us look at each of the changes.

```
! ------ Plotting window
! (x axis = 1.5 times y axis)
SET WINDOW -40,140,-10,110
! ------ Initial conditions
LET r = 35
LET i = 0
! ------ Calculational loop
D0
LET Theta = (2*Pi/100)*i
LET x = 50 + r*COS(Theta)
LET y = 50 + r*SIN(Theta)
PLOT x,y;
LET i = i+1
LOOP UNTIL i = 100
```

## END

**Figure 5** *Listing of the BASIC program.* 

## Variable Names

Our command

LET  $\theta = (2*\pi/100)*i$ 

has been rewritten in the form

LET Theta = (2\*Pi/100)\*i

Unfortunately, only a few special symbols are available in the font chosen by True BASIC. When we want a symbol like  $\theta$  and it is not available, we can spell it out as we have done.

We have spelled out the name "Pi" for  $\pi$ , because BASIC understands that the letters "Pi" stand for the numerical value of  $\pi$ . ("Pi" is what is called a reserved word in True BASIC.)

## **Multiplication**

We are used to writing an expression like

 $r\cos(\theta)$ 

and assuming that the variable r multiplies the function  $\cos(q)$ . In BASIC you must always use an "\*" for multiplication, thus the correct way to write  $r \cos(\theta)$  is

 $r*\cos(\theta)$ 

Similarly we had to write 2\*Pi rather than 2Pi in the line defining Theta.

## **Plotting a Point**

Our command

Plot a point at (x, y)

becomes in BASIC

PLOT x,y

It is not as descriptive as our command, but it works the same way.

#### **Comment Lines**

In a number of places in the BASIC program we have added lines that begin with an exclamation point "!". These are called "comment lines" and are included to make the program more readable. A comment line has no effect on the operation of the program. The computer ignores anything on a line following an exclamation point. Thus the two lines

LET i = i + 1LET i = i + 1 ! Increment i

are completely equivalent. (If you write a command that does something peculiar, you can explain it by adding a comment as we did above.)

## **Plotting Window**

The only really new thing in the BASIC program of Figure (5) is the **SET WINDOW** command. We are going to plot a number of points whose x and y values all fall within the range between 0 and 100. We have to tell the computer what kind of scale to use when plotting these points.

In the command

SET WINDOW -40, 140, -10, 110

the computer adjusts the plotting scales so that the computer screen starts at -40 and goes to +140 along the horizontal axis, and ranges from -10 to +110 along the vertical axis, as shown in Figure (6).



**Figure 6** Using the SET WINDOW command.

This setting gives us plenty of room to plot anything in the range 0 to 100 as shown by the dotted square in Figure (6).

When we are plotting a circle, we would like to have it look like a circle and not get stretched out into an ellipse. In other words we would like a horizontal line 10 units long to have the same length as a vertical line 10 units long. True BASIC for the Macintosh computer could have easily have done this because Macintosh pixels are square, so that equal horizontal and vertical distances should simply contain equal numbers of pixels. (A pixel is the smallest dot that can be drawn on the screen. A standard Macintosh pixel is 1/ 72 of an inch on a side, a dimension consistent with typography standards.)

However True BASIC also works with IBM computers where there is no standard pixel size or shape. To handle this lack of standardization, True BASIC left it up to the user to guess what choice in the SET WIN-DOW command will give equal x and y dimensions. This is an unfortunate compromise.

If you are using a Macintosh MacPlus, Classic or SE, one of the computers with the 9" screens, set the horizontal dimension 1.5 times bigger than the vertical one, use the full screen as an output window, and the dimensions will match (circles will be circular and squares square.) If you have any other screen or computer, you will have to keep adjusting the SET WIN-DOW command until you get the desired results. (Leave the y axis range from -10 to +110, and adjust the x axis range. For the 15" screen of the iMac, we got a round circle plot for x values from -33 to +133.)

## **Practice**

The best way to learn how to handle BASIC programs is to start with a working program like the one in Figure (5), and make small modifications and see what happens. Below are a series of exercises designed to give you this practice, while at the same time introducing some techniques that will be useful in the analysis of strobe photographs. When you finish these exercises, you will be ready to use BASIC as a tool for predicting the motion of projectiles, both without and with air resistance, which is the subject of the remainder of the chapter.

## Exercise 2 A Running Program

Get a copy of True BASIC (preferably version 2.0 or later), launch it, and type in the program shown in Figure (5). Type it in just as we have printed it, with the same indentations at the beginning of the lines, and the same comments. Then run the program. You should get an output window that has the circle of dots shown in Figure (7).

If something has gone wrong, and you do not get this output, first check that you have typed exactly what we printed in Figure 5. If that doesn't work, get help from a friend, advisor, computer center, whatever. Sometimes the hardest part of programming is turning on the equipment and getting things started properly.

Once you get your circle of dots, save a copy of the program.



**Figure 7** The circle of dots plotted by the program shown in Figure (5).

## Exercise 3 Plotting a Circular Line

It's pretty hard to see the dots in Figure (7). The output can be made more visible if lines are drawn connecting the dots to give us a circular line. In BASIC it is very easy to connect the dots you are plotting. You simply add a semicolon after the PLOT command. I.e., change the command

PLOT x, y ! Plots dots

to the command

PLOT x, y; ! Plots lines

The result is shown in Figure (8).

Modify your program by changing the PLOT command as shown, and see that your output looks like Figure (8).

(Optional—There is a short gap in the circle on the right hand side. Can you modify your program to eliminate this gap?)





#### Exercise 4 Labels and Axes

Although we have succeeded in drawing a circle, the output is fairly bare. It is impossible to tell, for example, that we have a circle of radius 35, centered at x = y = 50. We can get this information into the output by drawing axis and labeling them. This can be done by adding the following lines near the beginning of the program, just after the SET WINDOW command

ļ		––– Dr	naw	& lab	еl	axes
	BOX L	INES C	),10	10,0,1	00	
	PLOT	TEXT,	ΑT	-3,0	: "	0"
	PLOT	TEXT,	ΑT	-13,9	6:	"y=100"
	PLOT	TEXT,	AT	101,0	:	"×=100"

The results of adding these lines are shown in Figure (9). The BOX LINES command drew a box around the region of interest, and the three PLOT TEXT lines gave us the labels seen in the output.

Add the 5 lines shown above to your program and see that you get the results shown in Figure (9). Save a copy of that version of the program using a new name. Then find out how the BOX LINES and PLOT TEXT commands work by making some changes and seeing what happens.



#### Figure 9

A box, drawn by the BOX LINES command makes a good set of axes. You can then plot text where you want it.

#### Exercise 5a Numerical Output

Sometimes it is more useful to see the numerical results of a calculation than a plot. This can easily be done by replacing the PLOT command by a PRINT command.

To do this, go back to your original circle plotting program (the one shown in Figure (5) which we asked you to save), and change the line

to the two lines

What we have done is added the PRINT line, and then put an exclamation point at the beginning of the PLOT line so that the computer would ignore the PLOT command. (We left the PLOT line in so that we could use it later.) If we ran the program we get a whole bunch of printing, part of which is shown in Figure (10). Do this and see that you get the same results.

x	=	79.5515	y =	68.7539
x	=	78.3156	y =	70.5725
x	=	76.968	y =	72.3098
x	=	75.5139	y =	73.9591
x	=	73.9591	y =	75.5139
x	=	72.3098	y =	76.968
×	=	70.5725	y =	78.3156
x	=	68.7539	y =	79.5515
x	=	66.8614	y =	80.6707
×	=	64.9023	y =	81.6689
x	=	62.8844	y =	82.5422
x	=	60.8156	y =	83.287
x	=	58.7041	y =	83.9004
x	=	56.5583	y =	84.3801
x	=	54.3867	y =	84.724
х	=	52.1977	y =	84.9309
х	=	50.	y =	85
x	=	47.8023	y =	84.9309

#### Figure 10

If we print the coordinates of every point, we get too much output.

## **Selected Printing (MOD Command)**

The problem with the output in Figure (10) is that we print out the coordinates of every point, and we may not want that much information. It may be more convenient, for example, if we print the coordinates for every tenth point. To do this, we use the following trick. We replace the PRINT command

PRINT "x = ";x, "y = ";y

by the command

IF MOD(i,10) THEN PRINT "x =";x, "y = ";y

To understand what we did, remember that each time we go around the loop, the variable i is incriminated by 1. The first time i = 0, then it equals 1, then 2, etc.

The function MOD(), stands for the mathematical term "modulus". If we count modulus 3, for example, we count: 0, 1, 2, and then go back to zero when we hit 3. Comparing regular counting with counting MOD 3, we get:

regular counting:	0123456789
counting MOD 3:	0120120120

Counting MOD 10, we go:

0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 0, 1, 2, 3, ... etc. Every time we get up to a power of ten, we go back to zero.

The command MOD(i, 10) means evaluate the number i counting modulus 10. Thus when i gets to 10, MOD(i, 10) goes back to zero. When i gets to 20, MOD(i, 10) goes back to zero again. Thus as i increases, MOD(i, 10) goes back to zero every time i hits a power of 10.

In the command

IF MOD(i,10) THEN PRINT "x = ";x, "y = ";y no printing occurs until i increases to a power of ten. Then we do get a print. The result is that with this command the coordinates of every tenth point are printed, and there is no printing for the other points, as we see in Figure (11).

## Exercise 5b

Take your program from Exercise (5a), modify the print command with the MOD statement, and see that you get the results shown in Figure (11). Then figure out how to print every 5th point or every 20th point. See if it works.

x =	85	y =	50
× =	78.3156	y =	70.5725
× =	60.8156	y =	83.287
× =	39.1844	y =	83.287
× =	21.6844	y =	70.5725
× =	15	y =	50.
× =	21.6844	y =	29.4275
× =	39.1844	y =	16.713
× =	60.8156	y =	16.713
x =	78.3156	y =	29.4275

#### Figure 11

The coordinates of every tenth point is printed when we use the MOD command.

#### Exercise 6 Plotting Crosses

Our last exercise will be to have the computer plot both a circle, and a set of crosses located at every tenth point along the circle as shown in Figure (12). This is about as fancy a plot as we will need in the course, so that you are almost through practicing the needed fundamentals.

To plot the crosses seen in Figure (12) we added what is called a "subroutine" shown at the bottom of Figure (13).

To get the program shown in Figure (13), go back to the program of Exercise (3) (we asked you to save it), and add the command

## IF MOD(i, 10) = 0 THEN CALL CROSS

where "CROSS" is the name of the subroutine at the bottom of Figure (13). You can see that the IF MOD(i, 10) = 0 part of the command has the subroutine called at every tenth point.



**Figure 12** *Here we use the MOD command and a subroutine to plot a cross at every tenth dot.* 

Next add in the subroutine lines as shown in Figure (13) (your program should look just like Figure 13) and see if you get the results shown in Figure (12). When you have a running program, figure out how to make the crosses bigger or smaller. How can you plot twice as many crosses?

```
! ----- Plotting window
           (x axis = 1.5 times y axis)
1
   SET WINDOW -40,140,-10,110
! ----- Draw & label axes
  BOX LINES 0,100,0,100
   PLOT TEXT, AT -3,0 : "O"
   PLOT TEXT, AT -13,96: "x=100"
   PLOT TEXT, AT 101,0 : "y=100"
! ----- Initial conditions
  LET r = 35
  LET i = 0
! ----- Calculational loop
   DO
     LET Theta = (2*PI/100)*i
     LET x = 50 + r*COS(Theta)
     LET y = 50 + r*SIN(Theta)
     PLOT x,y
      IF MOD(i,5) = 0 THEN CALL CROSS
     LET i = i+1
  LOOP UNTIL i = 100
! ----- Subroutine "CROSS" draws a cross at x,y.
   SUB CROSS
      PLOT LINES: X-4,Y; X+4,Y
      PLOT LINES: X, Y-4; X, Y+4
   END SUB
```

END

#### Figure 13

The complete BASIC program for drawing the picture shown in Figure (12).

## PREDICTION OF MOTION

Now that we have the techniques to handle a repetitive calculation we can return to the problem of using the step-by-step method to predict the motion of a projectile. The idea is that we will convert our graphical analysis of strobe photographs, discussed in Chapter 3, into a pair of equations that predict the motion of the projectile one step at a time. We will then see how these equations can be applied repeatedly to predict motion over a long period of time.

Figure (14a) is essentially our old Figure (3-16) where we used a strobe photograph to define the velocity of the projectile in terms of the projectile's coordinate vectors  $\vec{R}_i$  and  $\vec{R}_{i+1}$ . The result was

$$\vec{v}_i = \frac{\vec{S}_i}{\Delta t} = \frac{\vec{R}_{i+1} - \vec{R}_i}{\Delta t}$$
(4)

If we multiply Equation (4) through by  $\Delta t$  and rearrange terms, we get

$$\vec{R}_{i+1} = \vec{R}_i + \vec{v}_i \Delta t \tag{5}$$

which is the vector equation pictured in Figure (14a).



#### Figure 14a

To predict the next position  $\vec{R}_{i+1}$  of the ball, we add the ball's displacement  $\vec{S}_i = \vec{v}_i \Delta t$  to the present position  $\vec{R}_i$ .

Equation (5) can be interpreted as an equation that predicts the projectile's new position  $\vec{R}_{i+1}$  in terms of the old position  $\vec{R}_i$ , the old velocity vector  $\vec{v}_i$ , and the time step  $\Delta t$ . To emphasize this predictive nature of Equation (5), let us rename  $\vec{R}_{i+1}$  the new vector  $\vec{R}_{new}$ , and the old vectors  $\vec{R}_i$  and  $\vec{v}_i$ , as  $\vec{R}_{old}$  and  $\vec{v}_{old}$ . With this renaming, the equation becomes

$$\vec{R}_{new} = \vec{R}_{old} + \vec{v}_{old} * \Delta t$$
 (6)

which is illustrated in Figure (14b).

Equation (6) predicts the new position of the ball using the old position and velocity vectors. To use Equation (6) over again to predict the next new position of the ball, we need updated values for  $\vec{R}$ and  $\vec{v}$ . We already have  $\vec{R}_{i+1}$  or  $\vec{R}_{new}$  for the updated coordinate vector; what we still need is an updated velocity vector  $\vec{v}_{i+1}$  or  $\vec{v}_{new}$ .



#### Figure 14b

So that we do not have to number every point in our calculation, we label the current position "old", and the next position "new". To obtain the updated velocity, we use Figure (3-17), drawn again as Figure (15a), where the acceleration vector  $\vec{a_i}$  was defined by the equation

$$\vec{a}_i = \frac{\vec{v}_{i+1} - \vec{v}_i}{\Delta t} \tag{7}$$

Multiplying through by  $\Delta t$  and rearranging terms, Equation (7) becomes

$$\vec{\mathbf{v}}_{i+1} = \vec{\mathbf{v}}_i + \vec{\mathbf{a}}_i \Delta t \tag{8}$$

which expresses the new velocity vector in terms of the old velocity  $\vec{v_i}$  and the old acceleration  $\vec{a_i}$ , as illustrated in Figure (15a). Changing the subscripts from i + 1 and i to "new" and "old" as before, we get

$$\vec{v}_{new} = \vec{v}_{old} + \vec{a}_{old} * \Delta t$$
(9)

as our basic equation for the projectile's new velocity.

We have now completed one step in our prediction of the motion of the projectile. We start with the old position and velocity vectors  $\vec{R}_{old}$  and  $\vec{v}_{old}$ , and used Equations (6) and (9) to get the new vectors  $\vec{R}_{new}$  and  $\vec{v}_{new}$ . To predict the next step in the motion, we change the names of  $\vec{R}_{new}$ ,  $\vec{v}_{new}$  to  $\vec{R}_{old}$  and  $\vec{v}_{old}$  and repeat Equations (6) and (9). As long as we know the acceleration vector  $\vec{a}_i$  at each step, we can predict the motion as far into the future as we want.



#### Figure 15a

Once we get to the "new" position, we will need the new velocity vector  $\vec{v}_{new}$  in order to predict the next new position.

There are two important criteria for using this stepby-step method of predicting motion described above. One is that we must have an efficient method to handle the repetitive calculations involved. That is where the computer comes in. The other is that we must know the acceleration at each step. In the case of projectile motion, where  $\vec{a}$  is constant, there is no problem. We can also handle projectile motion with air resistance if we can use formulas like

$$\vec{a}_{air} = -K\vec{v}$$
  
 $\vec{a} = \vec{g} + \vec{a}_{air}$ 

shown in Figure (3-31). To handle more general problems, we need a new method for determining the acceleration vector. That new method was devised by Isaac Newton and will be discussed in the chapter on Newtonian Mechanics. In this chapter we will focus on projectile motion with or without air resistance so that we know the acceleration vectors throughout the motion.



## Figure 15b

*Yhe value of*  $\vec{v}_{new}$  *is obtained from the definition of acceleration*  $\vec{A} = (\vec{v}_{new} - \vec{v}_{old}) / \Delta t$ .

## TIME STEP AND INITIAL CONDITIONS

Equation (6) and (9) are the basic components of our step-by-step process, but there are several details to be worked out before we have a practical program for predicting motion. Two of the important ones are the choice of a time step  $\Delta t$ , and the initial conditions that get the calculations started.

In our strobe photographs we generally used a time step  $\Delta t = .1$  second so that we could do effective graphical work. If we turn the strobe up and use a shorter time step, then the images are so close together, the arrows representing individual displacement vectors are so short, that we cannot accurately add or subtract them. Yet if we turn the strobe down and use a longer  $\Delta t$ , our analysis becomes too coarse to be accurate. The choice  $\Delta t = .1$  sec is a good compromise.

When we are doing numerical calculations, however, we are not limited by graphical techniques and can get more accurate results by using shorter time steps. We will see that for the analysis of our strobe photographs, time steps in the range of .01 second to .001 second work well. Much shorter time steps, like a millionth of a second, greatly increase the computing time required while not giving more accurate results. If we use ridiculously short time steps like a nanosecond, the computer must do so many calculations that the roundoff error in the computer calculations begins to accumulate and the answers get worse, not better. Just as with graphical work there is an optimal time step. (Later we will have some exercises where you try various time steps to see which give the best results.)



#### Figure 16

By using a very short time Step dt in our computer calculation, we will closely follow the continuous path shown by the dotted lines. Thus we should use the instantaneous velocity vector  $\vec{v}_0$ , rather than the strobe velocity  $\vec{v}_0$  as our initial velocity.



#### Figure 17

The displacement  $\vec{v}_0 \Delta t$  is just half the displacement  $(\vec{R}_1 - \vec{R}_{-1})$ . This is an exact result for projectile motion, and quite accurate for most strobe photographs.

When we use a short time step of .01 seconds or less for analyzing our projectile motion photographs, we are close to what we have called the instantaneous velocity illustrated in Figure (3-32). But, as shown in Figure (16), the instantaneous velocity  $\vec{v}_0$  and the strobe velocity  $\vec{v}_0$  are quite different if the strobe velocity was obtained from a strobe photograph using  $\Delta t = .1$ second. To use the computer to predict the motion we see in our strobe photographs, we need the initial position  $\vec{R}_0$  and the initial velocity  $\vec{v}_0$  as the start for our step-by-step calculation. If we are going to use a very short time step in our computer calculation, then our first velocity vector should be the instantaneous velocity  $\vec{v}_0$ , not the strobe velocity .

This does not present a serious problem, because back in Chapter 3, Figure (3-33) reproduced here as Figure (17), we showed a simple method for obtaining the ball's instantaneous velocity from a strobe photograph. We saw that the instantaneous velocity  $\vec{v}_0$  was the average of the previous and following strobe velocities  $\vec{v}_{-1}$  and  $\vec{v}_1$ :

$$\vec{\underline{v}}_0 = \frac{\vec{v}_{-1} + \vec{v}_1}{2} \tag{10}$$

where  $\vec{v}_{-1} = \vec{S}_{-1} / \Delta t$  and  $\vec{v}_0 = \vec{S}_0 / \Delta t$ .

However, the sum of the two displacement vectors  $(\vec{S}_{-1} + \vec{S}_0)$  is just the difference between the coordinate vectors  $\vec{R}_1$  and  $\vec{R}_{-1}$  as shown in Figure (17). Thus the instantaneous velocity of the ball at Position (0) in Figure (17) is given by the equation

$$\vec{\underline{v}}_{0} = \frac{\vec{R}_{1} - \vec{R}_{-1}}{\Delta t}$$
(11)

If we use Equation (11) as the formula for the initial velocity in our step-by-step calculation, we are starting with the instantaneous velocity at Position (0) and can use very short time intervals in the following steps.

To avoid confusing the longer strobe time step and the shorter computer time step that we will be using in the same calculation, we will give them two different names as follows. We will use  $\Delta t$  for the longer strobe time step, which is needed for calculating the initial instantaneous velocity, and the name **dt** for the short computer time step.

$$\Delta t$$
 = time between strobe flashes  
dt = computer time step (12)

This choice of names is more or less consistent with calculus, where  $\Delta t$  is a small but finite time interval and dt is infinitesimal.

## AN ENGLISH PROGRAM FOR PROJECTILE MOTION

We are now ready to write out a program for predicting the motion of a projectile. The first version will be what we call an "English" program -- one that we can easily read and understand. Once we have checked that the program does what we want it to do, we will see what modifications are necessary to translate the program into BASIC.

## **English Program**

! ------ Initial conditions  
LET 
$$\Delta t = .1$$
  
LET  $\vec{R}_{old} = \vec{R}_0$   
LET  $\vec{V}_{old} = (\vec{R}_1 - \vec{R}_{-1})/(2*\Delta t)$   
LET  $T_{old} = 0$ 

```
! ------ Calculational loop

DO

LET \vec{R}_{new} = \vec{R}_{old} + \vec{V}_{old}*dt

LET \vec{A} = \vec{g}

LET \vec{V}_{new} = \vec{V}_{old} + \vec{A}*dt

LET T_{new} = T_{old} + dt

PLOT \vec{R}

LOOP UNTIL T > 1
```

## Figure 18

The first version of the English projectile motion program is shown in Figure (18). This program is designed to predict the motion of the steel ball projectile shown in Figure (3-8) and used for the drawings seen in Figures (15) and (16).

In the program we begin with a statement of the initial conditions – the starting point for the analysis of the motion. In this photograph, the strobe time step is  $\Delta t = .1$  seconds, and we are beginning the calculations at the position labeled  $\vec{R}_0$  in Figure (16). The instantaneous velocity at that point is given by the formula  $\vec{v}_0 = (\vec{R}_1 - \vec{R}_{-1})/2\Delta t$  as shown in Figure (17). These results appear in the program in the lines

LET 
$$\Delta t = .1$$
  
LET  $\vec{R}_{old} = \vec{R}_0$   
LET  $\vec{V}_{old} = (\vec{R}_1 - \vec{R}_{-1})/(2*\Delta t)$ 

Our new thing we are going to do in this program is keep track of the time by including the variable T in our calculations. We begin by setting T = 0 in the initial conditions, and then increment the clock by a computer time step dt every time we go around the calculation loop. This way T will keep track of the elapsed time throughout the calculations. The clock is initialized by the command

LET  $T_{old} = 0$ 

The computer time step dt plays a significant role in the program because we will want to adjust dt so that each calculational step is short enough to give accurate results, but not so short to waste large amounts of computer time. We will start with the value dt = .01 seconds, as shown by the command

LET dt = .01

Later we will try different time steps to see if the results change or are stable.

The important part of the program is the calculational loop which is repeated again and again to give us the step-by-step calculations. The calculations begin with the command

LET 
$$\vec{R}_{new} = \vec{R}_{old} + \vec{V}_{old} * dt$$

which is the calculation pictured in Figure (14b). Here we are using the short computer time step dt so that  $\vec{R}_{new}$  will be the position of the ball dt seconds after it was at  $\vec{R}_{old}$ .

The next line

LET 
$$\vec{A} = \vec{g}$$

simply tells us that for this projectile motion the ball's acceleration has the constant value  $\vec{g}$ . (Later, when we predict projectile motion with air resistance, we change this line to include the acceleration produced by the air resistance.)

To calculate the new velocity vector, we use the command

LET 
$$\vec{V}_{new} = \vec{V}_{old} + \vec{A} * dt$$

which is pictured in Figure (15b). Again we are using the short computer time step dt rather than the longer strobe rate  $\Delta t$ .

The last two lines inside the calculational loop are

```
LET T_{new} = T_{old} + dt
PLOT \vec{R}
```

The first of these increments the clock so that T will keep track of the elapsed time. Then we plot a point at the position  $\vec{R}$  so that we can get a graph of the motion of the ball.

The calculational loop itself is bounded by the DO and LOOP UNTIL commands:

DO  
...  
...  
LET 
$$T_{new} = T_{old} + dt$$
  
...  
LOOP UNTIL T > 1

Remember that with a DO – UNTIL loop there is a test to see if the condition, here T > 1, is met. If T has not reached 1, we go back to the beginning of the loop and repeat the calculations. Because of the command LET  $T_{new} = T_{old} + dt$ , T increments by dt each time around. At some point T will get up to one, the condition will be met, and we leave the loop. At that point the program is finished. (We chose the condition T > 1 to stop the calculation because the projectile spends less than one second in the strobe photograph. Later we may use some other criterion to stop the calculation.)

## A BASIC PROGRAM FOR PROJECTILE MOTION

The program in Figure (18) is quite close to a BASIC program. We have the LET statements and the Do - LOOP commands that appeared in our working BASIC program back in Figure (5). The only problem is that BASIC unfortunately does not understand vector equations. In order to translate Figure (18) into a workable BASIC program, we have to convert all the vector equations into numerical equations.

To do this conversion, we write the vector equation out as three component equations as shown below.

$$\vec{A} = \vec{B} + \vec{C} \tag{13}$$

becomes

 $Ax = Bx + Cx \tag{14a}$ 

$$Ay = By + Cy \tag{14b}$$

$$Az = Bz + Cz \tag{14c}$$

We saw this decomposition of a vector equation into numerical or scalar equations in Chapter 2 on vectors and Chapter 4 on calculus. (It should have been in Chapter 2 but was accidently left out. It will be put in.) If the motion is in two dimensions, say in the x-y plane, then we only need the x and y component Equations (14a) and (14b).

Let us apply this rule to translate the vector LET statement

LET 
$$\vec{R}_{new} = \vec{R}_{old} + \vec{V}_{old} * dt$$
 (15)

into two numerical LET statements. If we use the notation

$$\vec{R} = (Rx, Ry)$$
;  $\vec{V} = (Vx, Vy)$ 

we get, dropping the subscripts "new" and "old",

$$Rx = Rx + Vx * dt$$
(16a)

$$Ry = Ry + Vy*dt$$
(16b)

We can drop the subscripts "new" and "old" because in carrying out the LET statement the computer must use the old values of Rx and Vx to evaluate the sum Rx + Vx\*dt, and this result which is the new value of Rx is stored in the memory cell labeled "Rx".

To translate the initial conditions, we used the experimental values of the ball's coordinates given in Figure (3-10), the steel ball projectile motion strobe photograph we have been using for all of our drawings. These coordinates are reproduced below in Figure (19).

Ball coordinates			
-1) (8.3, 79.3)			
0) (25.9, 89.9)			
1) (43.2, 90.2)			
2) (60.8, 80.5)			
3) (78.2, 60.2)			
4) (95.9, 30.2)			

#### Figure 19

Experimental coordinates of the steel ball projectile, from Figure (3-10).

Using the fact that  $\vec{R}_0 = (25.9, 89.9)$ , we can write the equation

LET 
$$\vec{R}_{old} = \vec{R}_0$$

as the two equations

LET 
$$Rx = 25.9$$
  
LET  $Ry = 89.9$ 

In a similar way we use the experimental values for  $\vec{R}_1$  and  $\vec{R}_{-1}$  to evaluate the initial value of  $\vec{V}_{old}$ .

In Figure (20) we have converted the vector LET statements into scalar ones to obtain a workable BASIC program. We have also included the vector statements to the right so that you can see that the English and BASIC programs are essentially the same. We also added the SET WINDOW command so that the output could be plotted.

In Figure (21), we show the output from the Basic program of Figure (20). It looks about as bad as Figure (7), the output from our first circle plotting program. In the following exercises we will add axes, plot points closer together, and plot crosses every tenth of a second. In addition, we will get numerical output that can be compared directly with the experimental values shown in Figure (19).
# **BASIC Program**

# **English Program**

! ----- Plotting window Į. (x axis = 1.5 times y axis) SET WINDOW -40,140,-10,110 ! ------ Initial conditions LET DeltaT = .1 LET Rx = 25.9LET Ry = 89.9LET  $U_x = (43.2 - 8.4)/(2*.1)$ LET Uy = (90.2 - 79.3)/(2\*.1) LET T = 0! ----- Computer Time Step LET dt = .01! ----- Calculational loop DO LET Rx = Rx + Ux\*dt LET Ry = Ry + Vy\*dt LET  $A_X = 0$ LET Ay = -980LET Ux = Ux + Ax\*dt LET Vy = Vy + Ay\*dt LET T = T + dtPLOT Rx, Ry LOOP UNTIL T > 1

Projectile Motion program in both BASIC and English.

! ------ Initial conditions LET  $\Delta t = .1$ LET  $\vec{R}_{old} = \vec{R}_0$ LET  $\vec{V}_{old} = (\vec{R}_1 - \vec{R}_{-1})/(2*\Delta t)$ LET  $T_{old} = 0$ 

! ------ Computer Time Step LET dt = .01 ! ------ Calculational loop DO LET  $\vec{R}_{new} = \vec{R}_{old} + \vec{V}_{old} * dt$ LET  $\vec{A} = \vec{g}$ LET  $\vec{V}_{new} = \vec{V}_{old} + \vec{A} * dt$ LET  $T_{new} = T_{old} + dt$ PLOT  $\vec{R}$ LOOP UNTIL T > 1

END

Figure 21

END

Figure 20

Output from the BASIC program in Figure (20). (Look closely for the dots.)

# Exercise 7

Start BASIC, type the BASIC projectile motion program shown in Figure 20, and run it. Keep fixing it up until it gives output that looks like that shown in Figure 21.

# Exercise 8 Changing the Time Step

Reduce the time step to dt = .001 seconds. The plot should become essentially a continuous line.

# Exercise 9 Numerical Output

Change the plot command to a print statement to see numerical output. You can do this by turning the PLOT command into a comment, and adding a PRINT command as shown below.

**!PLOT Rx,Ry** 

PRINT "Rx = ";Rx, "Ry = ";Ry

Just as in Exercise 5, you will get too much output when you run the program. If you have done Exercise 8, the coordinates of the ball will be printed every thousandth of a second. Yet from the strobe photograph, you have data for tenth second intervals. The next two exercises are designed to reduce the output.

# Exercise 10 Attempt to reduce output

Replace the PRINT command of exercise 9 by the command

IF MOD(T,.1) = 0 THEN PRINT "Rx = ";Rx, "Ry = ";Ry

The idea is to pull the same trick we used in reducing the output in Exercise 5, going from Figure 10 to Figure 11. In The above MOD statement, we would hope that we would get output every time T gets up to a multiple of 0.1.

Try the modification of the PRINT command using MOD(T,.1) as shown above. When you do you will not get any output. The MOD(T,.1) command does not work, because the MOD function generally works only with integers. We will fix the problem in the next exercise.

## Exercise 11 Reducing Numerical Output

Because the MOD function works reliably only with integers, we will introduce a counter variable i like we had in our circle plotting program.

First we must initialize i. We can do that at the same time we initialize dt as shown.

! ----- Computer Time Step and Counter LET dt = .01 LET i = 0

Then we will increment i by 1 each time we go around the calculational loop, using the now familiar command LET i = i+1. If we are using a time step dt = .001 then we have to go around the calculational loop 100 times to reach a time interval of .1 seconds. To do this, our print command should start with IF MOD(i,100) = 0...Thus, inside the calculational loop, the Print command of Exercise 9 should be replaced by

## LET i = i+1

IF MOD(i,100) = 0 THEN PRINT "RX = ";RX, "RY = ";RY

Make the changes shown above, run your program, and see that you get the output shown below in Figure 22. Compare these results with the experimental values shown in Figure 19.

Rх	=	43.35	Ry	=	90.499
Rх	=	60.8	Ry	=	81.298
Rх	=	78.25	Ry	=	62.297
Rх	=	95.7	Ry	=	33.496
Rх	=	113.15	Ry	=	-5.105
Rх	=	130.6	Ry	=	-53.506
Rх	=	148.05	Ry	=	-111.707
Rх	=	165.5	Ry	=	-179.708
Rх	=	182.95	Ry	=	-257.509
Rх	=	200.4	Ry	=	-345.11

#### Figure 22

Numerical output from the projectile motion program, printed at time intervals of .1 seconds. These predicted results should be compared with the experimental results seen in Figure 3-10.

## 5-21

#### Exercise 12 Plotting Crosses

Now we have the MOD statement to reduce the printing output, we can use the same trick to plot crosses in the output at .1 second intervals. All we have to do is restore the PLOT command, change the MOD statement to

## IF MOD(i,100) = 0 THEN CALL CROSS

and add a cross plotting subroutine which should now look like

! ------ Subroutine "CROSS" draws a cross at Rx,Ry.

SUB CROSS

PLOT LINES: Rx-2,Ry; Rx+2,Ry

PLOT LINES: Rx,Ry-2; Rx,Ry+2

### END SUB

The only change from the CROSS subroutine in the circle plotting program is that the cross is now centered at coordinates (Rx,Ry) rather than (x,y) as before.

The complete cross plotting is shown in Figure (23), and the results are plotted in Figure (24). Modify your projectile motion program to match Figure (23), and see that you get the same results. (How did we stop the plotting outside the square box?)



**Figure 24** *Output from our BASIC projectile motion program of Figure 23.* 

#### **Projectile Motion Program**

```
! ----- Plotting window
ŗ
            (x axis = 1.5 times y axis)
   SET WINDOW -40,140,-10,110
! ----- Draw & label axes
   BOX LINES 0,100,0,100
   PLOT TEXT, AT -3,0 : "O"
  PLOT TEXT, AT -13,96: "y=100"
  PLOT TEXT, AT 101,0 : "x=100"
! ----- Initial conditions
  LET DeltaT = .1
  LET R_{x} = 25.9
  LET Ry = 89.9
  LET Ux = (43.2 - 8.3)/(2*.1)
  LET Uy = (90.2 - 79.3)/(2*.1)
  LET T = 0
  CALL CROSS
! ----- Computer Time Step
  LET dt = .001
  LET i = 0
! ----- Calculational loop
  DO
      LET Rx = Rx + Vx*dt
      LET Ry = Ry + Vy*dt
      LET A \times = 0
      LET Ay = -980
      LET Ux = Ux + Ax*dt
      LET Vy = Vy + Ay*dt
      LET T = T + dt
      LET i = i+1
      IF MOD(i,100) = 0 THEN CALL CROSS
      PLOT Rx, Ry
   LOOP UNTIL RX > 100
! ----- Subroutine "CROSS" draws
          ! a cross at Rx,Ry.
   SUB CROSS
       PLOT LINES: Rx-2, Ry; Rx+2, Ry
       PLOT LINES: Rx,Ry-2; Rx,Ry+2
   END SUB
END
```

### Figure 23

Projectile motion program that plots crosses every tenth of a second.

# PROJECTILE MOTION WITH AIR RESISTANCE

Projectile motion is an example of a very special kind of motion where the acceleration vector is constant – does not change in either magnitude or direction. In this special case we can easily use calculus to predict motion far into the future.

But let the acceleration vector change even by a small amount, as in the case of projectile motion with air resistance, and a calculus solution becomes difficult or impossible to obtain. This illustrates the important role the acceleration vector plays in the prediction of motion, but overemphasizes the importance of motion with constant acceleration.

With a computer solution, very little additional effort is required to include the effects of air resistance. We will be able to adjust the acceleration for different amounts or kinds of air resistance. The point is to develop an



## Figure 25a

The styrofoam projectile of Figure (3-30a). We have printed a negative of the photograph to show the grid lines more distinctly.

intuition for the role played by the acceleration vector. We will see that *if we know a particle's acceleration, have a formula for it, and know how the particle started moving, we can predict where the particle will be at any time in the future.* 

Once we have gained experience with this kind of prediction, we can then focus our attention on the core problem in mechanics, namely finding a general method for determining the acceleration vector. As we mentioned, the general method was discovered by Newton and will be discussed shortly in the chapter on Newtonian Mechanics.

In our study of the effects of air resistance, we will use as our main example the styrofoam ball projectile shown in Figures (3-30a, b) and reproduced here as Figures (25a, b). To obtain the coordinates listed in Figure (25b), each image was enlarged and studied separately. As a result, these coordinates should be accurate to within half a millimeter (except for possible errors due to parallax in taking the photograph).



## Figure 25b

To obtain as accurate a value as we could for each ball coordinate, each image was enlarged and studied separately. Figure (26), a reproduction of Figure (3-31), is a detailed analysis of the ball's acceleration at Position (3). As shown in Figure (26) we can write the formula for the ball's acceleration vector  $\vec{A}$  in the form

$$\vec{A} = \vec{g} + \vec{A}_{air}$$
(17)

where one possible formula for  $\vec{A}_{air}$  is

$$\vec{A}_{air} = -K\vec{V}$$
(18)

 $\vec{V}$  being the instantaneous velocity of the ball.

In Equation (17),  $\vec{A}_{air}$  is defined as the change from the normal acceleration  $\vec{g}$  the projectile would have without air resistance. As we see,  $\vec{A}_{air}$  points opposite to  $\vec{V}$ , which is the direction of the wind we would feel if we were riding on the ball. Figure (26) suggests the physical interpretation that this wind is in effect blowing the acceleration vector back. It suggests that acceleration vectors can be pushed or pulled around, which is the underlying idea of Newtonian mechanics. In Figure (26) the earth is pulling down on the ball which gives rise to the component  $\vec{g}$  of the ball's acceleration, and the wind is pushing back to give rise to the component  $\vec{A}_{air}$ .



The simplest formula we can write which has  $\vec{A}_{air}$  pointing in the  $-\vec{V}$  direction is Equation (18),  $\vec{A}_{air} = -\vec{KV}$ , where K is a constant that we have to find from the experiment. If some choice of the constant K allows us to accurately predict all the experimental points in Figure (25), then we will have verified that Equation (18) is a reasonably accurate description of the effects of air resistance.

It may happen, however, that one choice of K will lead to an accurate prediction of one position of the ball, while another choice leads to an accurate prediction of another point, but no value of K gives an accurate prediction of all the points. If this happens, equation (18) may be inadequate, and we may need a more complex formula.

The next level of complexity is that K itself depends on the speed of the ball. Then  $\vec{A}_{air}$  would have a magnitude related to V<sup>2</sup>, V<sup>3</sup>, or something worse. In this case the air resistance is "nonlinear" and exact calculus solutions are not possible. But, as we see in Exercise 15, we can still try out different computer solutions.

In reality, when a sphere moves through a fluid like air or water, the resistance of the fluid can become very complex. At high enough speeds, the sphere can start shedding vortices, the fluid can become turbulent, and the acceleration produced by the fluid may no longer be directed opposite to the instantaneous velocity of the sphere. In Exercise 13 we take a close look at  $\vec{A}_{air}$  for all interior positions for the projectile motion shown in Figure (25). We find that to within experimental accuracy, for our styrofoam projectile  $\vec{A}_{air}$  does point in the  $-\vec{V}$  direction. Thus a formula like Equation (18) is a good starting point. We can also tell from the experimental data whether K is constant and what a good average value for K should be.

**Figure 26** *The air resistance is caused by the wind you would feel if you were riding on the ball.* 

# Air Resistance Program

Figure (20) was our BASIC program for projectile motion. We would now like to modify that program so that we can predict the motion of the Styrofoam ball shown in Figure (25). To do this, we must change the command

LET 
$$\vec{A} = \vec{g}$$

to the new command

LET 
$$\vec{A} = \vec{g} - \vec{KV}$$
 (19)

and try different values for K until we get the best agreement between prediction and experiment.

A complete program with this modification is shown in Figure (27). In this program we see that Equation (19) has been translated into the two component equations

LET 
$$Ax = 0 - K*Vx$$
  
LET  $Ay = -980 - K*Vy$ 

In addition, we are printing numerical output at .1 sec intervals so that we can accurately compare the predicted results with the experimental ones. In the line

LET K = ...

which appears in the Initial Conditions, we are to plug in various values of K until we get the best agreement that we can between theory and experiment.

Finding K does not have to be complete guesswork. In Exercise 13 we ask you to do a graphical analysis of the Styrofoam ball's acceleration at several positions using the enlargements provided. From these results you should choose some best average value for K and use that as your initial guess for K in your computer program. Then fine tune K until you get the best agreement you can. We ask you to do this in Exercise 14.

Once you have a working program that predicts the motion of the Styrofoam ball in Figure (25), you can easily do simulations of different strengths of air resistance. What if you had a steel ball being projected through a viscous liquid like honey? The viscous liquid might have the same effect as air, except that the

resistance constant K should be much larger. With the computer, you can simply use larger and larger values of K to see the effects of increasing the air or fluid viscosity. We ask you to do this in Exercise 15. This is a very worthwhile exercise, for as the fluid viscosity increases, as you increase K, you get an entirely new kind of motion. There is a change in the qualitative character of the motion which you can observe by rerunning the program with different values of K.

```
! ----- Plotting window
   (x axis = 1.5 times y axis)
ļ.
  SET WINDOW -40,140,-10,110
! ----- Initial conditions
  LET DeltaT = .1
  LET Rx = ...
                                Use initial values
  LET Ry = ...
  LET Ux = ...
                                from Figure (25).
  LET Vy = \dots
  LET T = 0
                                Try different
  LET K = ...
                                values of K
! ----- Computer Time Step
  LET dt = .001
  LET i = 0
! ----- Calculational loop
  DO
     LET Rx = Rx + Vx*dt
     LET Ry = Ry + Vy*dt
     LET A_{X} = 0
                 – K*V×
                                New formula for A
     LET Ay = -980 - K*Vy
     LET Ux = Ux + Ax*dt
     LET Vy = Vy + Ay*dt
     LET T = T + dt
     LET i = i+1
     IF MOD(i,100) = 0 THEN PRINT "Rx = ";Rx, "Ry = ";Ry
  LOOP UNTIL T > 1
  END
```

# Figure 27

BASIC program for projectile motion with air resistance. It is left to the reader to insert appropriate initial conditions, and choose values of the air resistance constant K. In Exercise 16, we show you one way to modify the air resistance formulas to include nonlinear effects, i.e., to allow  $\vec{A}_{air}$  to depend on V<sup>2</sup> as well as V. What we do is first use the Pythagorean theorem to calculate the magnitude V of the ball's speed and then use that in a more general formula for  $\vec{A}_{air}$ . The English lines for this are

LET 
$$V = \sqrt{V_x^2 + V_y^2}$$
  
LET  $\vec{A} = \vec{g} - K(1 + K2 * V)\vec{V}$  (20)

where we now try to find values of K and K2 that improve the agreement between prediction and experiment. The translation of these lines into BASIC is shown in Exercise 16.

#### **Exercise 13 Graphical Analysis**

Figures (28 a,b,c,d,e) are accurate enlargements of sections of Figure (25b). In each case we show three positions of the Styrofoam projectile so that you can determine the ball's instantaneous velocity  $\vec{V}$  at the center position. Using the section of grid you can determine the magnitude of both  $\vec{V} \Delta t$  and  $\vec{A}_{air} \Delta t^2$ . From that, and the fact that  $\Delta t = .1$  sec, you can then determine the size of the air resistance constant K using the equation  $\vec{A}_{air} = -K\vec{V}$ .

Do this for each of the diagrams, positions 0 through 4 and then find a reasonable average value of K. How constant is K? Do you have any explanation for changes in K?



**Figure 28a** Blowup of position 0 in Figure 25b.



Figure 28b

Blowup of position 1 in Figure 25b.



Figure 28c

Blowup of position 2 in Figure 25b.

# **Exercise 14 Computer Prediction**

Starting with the Basic program shown in Figure (27) use the experimental values shown in Figure (25b), reproduced below, to determine the initial conditions for the motion of the ball. Then use your best value of K from Exercise 13 as your initial value of K in the program. By trial and error, find what you consider the best value of K to bring the predicted coordinates into reasonable agreement with experiment.

-1) (	5.35,	94.84)
0) (2	24.03,1	101.29)
1) (4	10.90,	97.68)
2) (5	56.52,	85.15)
3) (7	0.77,	64.56)
4) (8	33.48,	36.98)
5) (9	95.18,	3.86)

## Exercise 15 Viscous Fluid

After you get your program of Exercise 14 working, allow the program to print out numerical values for up to T = 15 seconds. After about 10 seconds, the nature of the motion is very different than it was at the beginning. Explain the difference. (You may be able to see the difference better by printing Vx and Vy rather than Rx and Ry.)

You will see the same phenomenon much faster if you greatly increase the air resistance constant K. Redo your program to plot the output, drawing crosses every .1 seconds. Then rerun the program for ever increasing values of K. Explain what you see.



Figure 28d Blowup of position 3 in Figure 25b.

Ò

Blowup of position 4 in Figure 25b.

#### Exercise 16 Nonlinear Air Resistance (optional)

In Exercise 14, you probably found that you were not able to precisely predict all the ball positions using one value of K. In this exercise, you allow K to depend on the ball's speed v in order to try to get a more accurate prediction. One possibility is to use the following formulas for  $\vec{A}_{air}$ , which we mentioned earlier:

LET 
$$V = \sqrt{V_x^2 + V_y^2}$$
  
LET  $\vec{A} = \vec{g} - K(1 + K2*V)\vec{V}$  (20)

With Equations (20), you can now adjust both K and K2 to get a better prediction. These equations are translated into BASIC as follows.

LET V = SQR(Vx\*Vx + Vy\*Vy) LET Ax = 0 - K\*(1 + K2\*V)\*VxLET Ay = -980 - K\*(1 + K2\*V)\*Vy

Make these modifications in the program of Exercise 14, and see if you can detect evidence for some  $V^2$  dependence in the air resistance.



**Figure 29** Additional wind created by fan.

#### Exercise 17 Fan Added

In Figure (30), on the next page, we show the results of placing a rack of small fans to the right of the styrofoam ball's trajectory in order to increase the effect of air resistance. Now, someone riding with the ball should feel not only the wind due to the motion of the ball, but also the wind of the fans, as shown in Figure (29).

Our old air resistance formula

LET 
$$\vec{A} = \vec{g} + K(-\vec{V}ball)$$

should probably be replaced by a command like

LET 
$$\vec{A} = \vec{g} + K(-\vec{V}ball + \vec{V}fan)$$

Translated into BASIC, this would become

LET 
$$Ax = 0 + K*(-Vx - Vfan)$$
  
LET  $Ay = -980 + K*(-Vy + 0)$  (21)

where  $\mathbf{V}$ ball = (Vx,Vy) is the current velocity of the ball, and  $\mathbf{V}$ fan = (-Vfan,0) is the wind caused by the fan. We assume that this wind is aimed in the -x direction and has a magnitude Vfan. We now have two unknown parameters K and Vfan which we can adjust to match the experimental results shown in Figure (30).

Do this, starting with the value of K that you got from the analysis of the styrofoam projectile in Figure 25b (Exercise 13 or 14). Does your resulting value for Vfan seem reasonable? Can you detect any systematic error in your analysis? For example, should Vfan be stronger near the fans, and get weaker as you move left? COMPOSITE STROBE PHOTOGRAPH



#### Figure 30

Styrofoam projectile with a bank of fans. In order to get more air resistance, we added a bank of small fans as shown. This Strobe "photograph" was taken with the Apple II Strobe system.



#### Figure 31

In this diagram, the Apple II computer has calculated and plotted the centers of each of the images seen in the composite strobe photograph on the left.

	DELTA T =	3/30 SECOND	(TIME BETWEEN D	ROSSES)	
CROSS#1	X = 5.3	Y = 46.2	CROSS#2	X = 20.3	Y = 62.1
CROSS#3	X = 32.9	Y = 67.7	CROSS#4	X = 43.4	Y = 64.3
CROSS#5	X = 52.2	Y = 51.6	CROSS#6	X = 59.1	Y = 30.3
	INITIAL I	MAGES			
	IMAGE#1	X = 5.3	Y = 46.2		
	IMAGE#2	X = 10.8	Y = 52.8		
	IMAGE#3	X = 15.8	Y = 58.1		
	IMAGE#4	X = 20.3	Y = 62.1		
	IMAGE#5	X = 24.8	Y = 65.1		
	IMAGE#6	X = 28.8	Y = 66.9		
	IMAGE#7	X = 32.9	Y = 67.7		

#### Figure 32

The Apple II also prints out the Coordinates of each image. The time  $\Delta t$  between crosses is 1/10 sec. Between the dots there is a 1/30 sec time interval. The coordinates of the initial 7 dots are printed to help determine the initial instantaneous velocity of the ball.

# Chapter 6 Mass

By now we have learned how to use either calculus or the computer to predict the motion of an object whose acceleration is known. But in most problems we do not know the acceleration, at least initially. Instead we may know the forces acting on the object, or something about the object's energy, and use this information to predict motion. This approach, which is the heart of the subject of mechanics, involves mass, a concept which we introduce in this chapter.

In the metric system, mass is measured in grams or kilograms, quantities that should be quite familiar to the reader. It may be surprising that we devote an entire chapter to something that is measured daily by grocery store clerks in every country in the world. But the concept of mass plays a key role in the subject of mechanics. Here we focus on developing an experimental definition of mass, a definition that we can use without modification throughout our discussion of physics.

After introducing the experimental definition, we will go through several experiments to determine how mass, as we defined it, behaves. In low speed experiments, the kind we can do using air tracks in demonstration lectures, the results are straightforward and are what one expects. But when we consider what would happen if similar experiments were carried out with one of the objects moving at speeds near the speed of light, we predict a very different behavior for mass. This new behavior is summarized by the Einstein mass formula, a strikingly simple result that one might guess, but which we cannot quite derive from the definition of mass, and the principle of relativity alone. What is needed in addition is the law of conservation of linear momentum which we will discuss in the next chapter.

One of the striking features of Einstein's special theory of relativity is the fact that nothing, not even information, can travel faster than the speed of light. We can think of nature as having a speed limit c. In our world, speed limits are hard to enforce. We will see that the Einstein mass formula provides nature with an automatic way of enforcing its speed limit.

Einstein's mass formula appears to predict that no particle can quite reach the speed of light. We end the chapter with a discussion of how to handle particles, like photons and possibly neutrinos, that do travel at precisely the speed of light.

# **DEFINITION OF MASS**

In everyday conversation the words *mass* and *weight* are used interchangeably. Physicists use the words mass and weight for two different concepts. Briefly, we can say that the weight of an object is the force that the object exerts against the ground, and we can measure weight with a device such as a bathroom scale. The weight of an object can change in different circumstances. For example, an astronaut who weighs 180 pounds while standing on the ground, floats freely in an orbiting space capsule. If he stood on a bathroom scale in an orbiting space craft, the reading would be zero, and we would say he is weightless. On the other hand the mass of the astronaut is the same whether he is in orbit or standing on the ground. An astronaut in orbit does not become massless. Mass is not what you measure when you stand on the bathroom scales.

What then is mass? One definition, found in the dictionary, describes mass as the property of a body that is a measure of the amount of material it contains. Another definition, which is closer to the one we will use, says that the more massive an object, the harder it is to budge.

Both of these definitions are too vague to tell us how to actually measure mass. In this section we will describe an experimental definition of mass, one that provides



# **Recoil Experiments**

As a crude experiment suppose that the two skaters shown in Figure (6-1), a father and a child, stand in front of each other at rest and then push each other apart. The father hardly moves, while the child goes flying off. The father is more massive, harder to budge. No matter how hard or gently the skaters push apart, the big one always recoils more slowly than the smaller one. We will use this observation to define mass.

In a similar but more controlled experiment, we replace the skaters by two carts on what is called an *air track*. An air track consists of a long square metal tube with a series of small holes drilled on two sides as shown in Figure (6-2). A vacuum cleaner run backwards blows air into the tube, and the air escapes out through the small holes. The air carts have V-shaped bottoms which ride on a thin film of air, allowing the carts to move almost without friction along the track.

To represent the two skaters pushing apart on nearly frictionless ice, we set up two carts with a spring between them as shown in Figure (6-3a). A thread is tied between the carts to keep the spring compressed. When we burn the thread, the carts fly apart as shown



## Figure 1

Two skaters, a father and a son, standing at rest on frictionless ice, push away from each other. The smaller, less massive child recoils faster than the more massive father.



#### Figure 2

End view of an air track. Pressurized air from the back side of a vacuum cleaner is fed into a square hollow metal tube, and flows out through a series of small holes. A cart, riding on a film of air, can move essentially without friction along the track. in Figure (6-3b). If the two carts are made of similar material, but one is bigger than the other, the big one will recoil at lesser speed than the small one. We say that the big cart, the one that comes out more slowly, has more mass than the small one.

Because we can precisely measure the speeds  $v_A$  and  $v_B$  of the recoiling air carts, we can use the experiment pictured in Figures (6-3a,b) to define the mass of the carts. Let us call  $m_A$  and  $m_B$  the masses of carts A and B respectively. The simplest formula relating the masses of the carts to the recoil speeds, a formula that has the more massive cart recoiling at less speed is

$$\frac{\mathbf{m}_{A}}{\mathbf{m}_{B}} = \frac{\mathbf{v}_{B}}{\mathbf{v}_{A}} \quad recoil \ definition \ of \ mass \tag{1}$$

In words, Equation 1 says that the ratio of the masses is inversely proportional to the recoil speeds. I.e., if  $m_A$  is the small mass, the  $v_B$  is the small speed.



#### Figure 3

Recoil experiment. To simulate the two skaters pushing apart, we place two carts on an air track with a compressed spring between them. The carts are held together by a string. When the string is burned, the carts fly apart as did the skaters. The more massive cart recoils at a smaller speed  $(v_B < v_A)$ .

## **Properties of Mass**

Since we now have an explicit prescription for measuring mass, we should carry out some experiments to see if this definition makes sense. Our first test is to see if the mass ratio  $m_A / m_B$  changes if we use different strength springs in the recoil experiment. If the ratio of recoil speeds  $v_B / v_A$ , and therefore the mass ratio, depends upon what kind of spring we use, then our definition of mass may not be particularly useful.

In the appendix to this chapter, we describe apparatus that allows us to measure the recoil speeds of the carts with fair precision. To within an experimental accuracy of 5% to 10% we find that the ratio  $v_B / v_A$  of the recoil speeds does not depend upon how hard the spring pushes the carts apart. When we use a stronger spring, both carts come out faster, in such a way that the speed ratio is unchanged. Thus to the accuracy of this experiment we conclude that the mass ratio does not depend upon the strength of the spring used.

## Standard Mass

So far we have talked about the ratio of the masses of the two carts. What can we say about the individual masses  $m_A$  or  $m_B$  alone? There is a simple way to discuss the masses individually. What we do is select one of the masses, for instance  $m_B$ , as the *standard mass*, and measure all other masses in terms of  $m_B$ . To express  $m_A$  in terms of the standard mass  $m_B$ , we multiply both sides of Equation (1) through by  $m_B$  to get

$$m_{A} = m_{B} \left( \frac{v_{B}}{v_{A}} \right) \begin{array}{c} \textit{formula for } m_{A} \\ \textit{in terms of the} \\ \textit{standard mass } m_{B} \end{array}$$
(2)

For a standard mass, the world accepts that the platinum cylinder kept by the International Bureau of Weights and Measures near Paris, France, is precisely one kilogram. If we reshaped this cylinder into an air cart and used it for our standard mass, then we would have the following explicit formula for the mass of cart A recoiled from the standard mass.

$$m_{A} = (1 \text{ kilogram}) \times \left(\frac{V_{std}}{V_{A}}\right) \begin{array}{l} \text{using the} \\ \text{one kilogram} \\ \text{cylinder for our} \\ \text{standard mass} \end{array}$$
(3)

where  $V_{std}$  is the recoil speed of the standard mass. Once we have determined the mass of one of our own carts, using the standard mass and Equation (3), we can then use that cart as our standard and return the platinum cylinder to the French.

Of course the French will not let just anybody use their standard kilogram mass. What they did was to make accurate copies of the standard mass, and these copies are kept in individual countries, one of them by the National Institute of Standards and Technology in Washington, DC which then makes copies for others in the United States to use.

# Addition of Mass

Consider another experiment that can be performed using air carts. Suppose we have our standard cart of mass  $m_B$ , and two other carts which we will call C and D. Let us first recoil carts C and D from our standard mass  $m_B$ , and determine that C and D have masses  $m_C$  and  $m_D$  given by

$$m_{C} = m_{B} \left( \frac{v_{B}}{v_{C}} \right)$$
 ;  $m_{D} = m_{B} \left( \frac{v_{B}}{v_{D}} \right)$ 

Now what happens if, as shown in Figure (6-4), we tie carts C and D together and recoil them from cart B. How is the mass  $m_{C+D}$  of the combination of the two



## Figure 4

Addition of mass. If we tie two carts C and D together and recoil the pair from our standard mass  $m_A$ , and use the formula

$$m_{C+D} = m_B \left( \frac{v_B}{v_{C+D}} \right)$$

for the combined mass  $m_{C+D}$ , we find from experiment that  $m_{C+D} = m_C + m_D$ . In other words the mass of the pair of carts is the sum of the masses of the individual carts, or we can say that mass adds. carts related to the individual masses  $m_C$  and  $m_D$ ? If we perform the experiment shown in Figure (6-4), we find that

$$m_{C+D} = m_C + m_D \quad mass \ adds \tag{4}$$

The experimental result, shown in Equation (4), is that mass adds. The mass of the two carts recoiled together is the sum of the masses of the individual carts. This is the reason we can associate the concept of mass with the quantity of matter. If, for example, we have two identical carts, then together the two carts have twice as much matter and twice as much mass.

## Exercise 1

In physics labs, one often finds a set of brass cylinders of various sizes, each cylinder with a number stamped on it, representing its mass in grams. The set usually includes a 50-gm, 100-gm, 200-gm, 500-gm, and 1000-gm cylinder. Suppose that you were given a rod of brass and a hacksaw; describe in detail how you would construct a set of these standard masses. At your disposal you have a frictionless air track, two carts of unknown mass that ride on the track, the standard 1000-gm mass from France (which can be placed on one of the carts), and various things like springs, thread, and matches.

# **A Simpler Way to Measure Mass**

The preceding problem illustrates two things. One is that with an air track, carts, and a standard mass, we can use our recoil definition to measure the mass of an object. The second is that the procedure is clumsy and rather involved. What we need is a simpler way to measure mass.

The simpler way involves the use of a balance, which is a device with a rod on a pivot and two pans suspended from the rod, as shown in Figure (6-5). If the balance is properly adjusted, we find from experiment that if equal masses are placed in each pan, the rod remains balanced and level. This means that if we place an unknown mass in one pan, and add brass cylinders of known mass to the other pan until the rod becomes balanced, the object and the group of cylinders have the same mass. To determine the mass of the object, all we have to do is add up the masses of the individual cylinders.

# **Inertial and Gravitational Mass**

The pan balance of Figure (6-5) is actually comparing the downward gravitational force on the contents of the two pans. If the gravitational forces are equal, then the rod remains balanced. What we are noting is that there are equal gravitational forces on equal masses. This is an experimental result, not an obvious conclusion. For example, we could construct two air carts, one from wood and one from platinum. Keep adjusting the size of the carts until their recoil speeds are equal, i.e., until they have equal recoil masses. Then put these carts on the pan balance of Figure (6-5). Although the wood cart has a much bigger volume than the platinum one, we will find that the two carts still balance. The gravitational force on the two carts will be the same despite their large difference in size.

In 1922, the Swedish physicist Etvös did some very careful experiments, checking whether two objects, which had the same mass from a recoil type of experiment would experience the same gravitational force as measured by a pan balance type of experiment. He demonstrated that we would get the same result to one part in a billion. In 1960, R. H. Dicke improved Etvös' experiments to an accuracy of 1 part in 10<sup>11</sup>.



#### Figure 5

Schematic drawing of a pan balance. If the balance is correctly adjusted and if equal masses are placed in the pans, the rod will remain level. This allows us to determine an unknown mass simply by comparing it to a known one. It is common terminology to call what we measure in a recoil experiment the *inertial mass* of the object, and what we measure using a pan balance the *gravitational mass*. The experiments of Etvös and Dicke demonstrate that inertial mass and gravitational mass are equivalent to each other to one part in 10<sup>11</sup>. Is this a coincidence, or is there some fundamental reason why these two definitions of mass turn out to be equivalent? Einstein addressed this question in his formulation of a relativistic theory of gravity known as Einstein's *General Theory of Relativity*. We will have more to say about that later.

## Mass of a Moving Object

One reason we chose the recoil experiment of Figure (3) as our experimental definition of mass is that it allows us to study the mass of moving objects, something that is not possible with a pan balance.

From the air track experiments we have discussed so far, we have found two results. One is that the ratio of the recoil speeds, and therefore the ratio of the masses of the two objects, does not depend upon the strength of the spring or the individual speeds  $v_A$  and  $v_B$ . If we use a stronger spring so that  $m_A$  emerges twice as fast,  $m_B$  also emerges twice as fast so that the ratio  $m_A/m_B$  is unchanged.

In addition, we found that mass adds. If carts C and D have masses  $m_C$  and  $m_D$  when recoiled individually from cart B, then they have a combined mass  $m_{C,D} = m_C + m_D$  when they are tied together and both recoiled from cart B.

# **RELATIVISTIC MASS**

In our air track experiments, we found that the ratio of the recoil speeds did not depend upon the strength of the spring we used. However, when the recoil speeds approach the speed of light, this simple result can no longer apply. Because of nature's speed limit c, the ratio of the recoil speeds must in general change with speed.

To see why the recoil speed ratio must change, imagine an experiment involving the recoil of two objects of very different size, for example a bullet being fired from a gun as shown in Figure (6). Suppose, in an initial experiment not much gunpowder is used and the bullet comes out at a speed of 100 meters per second and the gun recoils at a speed of 10 cm/sec = .1 m/sec. For this case the speed ratio is 1000 to 1 and we say that the gun is 1000 times as massive as the bullet.

In a second experiment we use more gun powder and the bullet emerges 10 times faster, at a speed of 1000 meters per second. If the ratio of 1000 to 1 is maintained, then we predict that the gun should recoil at a speed of 1 meter per second. If we did the experiment, the prediction would be true.

But, as a thought experiment, imagine we used such powerful gun powder that the gun recoiled at 1% the speed of light. If the speed ratio remained at 1000 to 1, we would predict that the bullet would emerge at a speed 10 times the speed of light, an impossible result. The bullet cannot travel faster than the speed of light, the speed ratio cannot be greater than 100 to 1, and thus the ratio of the masses of the two objects must have changed.



## Figure 6

To discuss higher speed recoils, consider a bullet being fired from a gun. We are all aware that the bullet emerges at a high speed, but the gun itself also recoils. (The recoil of the gun becomes obvious the first time you fire a shotgun.) In this setup, the gunpowder is analogous to the spring, and the gun and bullet are analogous to the two carts. In the next section we will discuss experiments in which, instead of a bullet being fired by a gun, an electron is ejected by an atomic nucleus. The electron is such a small particle that it is often ejected at speeds approaching the speed of light. The nuclei we will consider are so much more massive that they recoil at low speeds familiar to us, speeds like that of a jet plane or earth satellite. At these low speeds the mass of an object does not change noticeably with speed. Thus in these electron recoil experiments, the mass of the nuclei is not changing due to its motion. Any change in the ratio of recoil speeds is due to a change in the mass of the electron as the speed of the electron approaches the speed of light.

We will see that as we push harder and harder on the electron, trying to make it go faster than the speed of light, *the mass of the electron increases instead*. It is precisely this increase in mass that prevents the electron emerging at a speed greater than the speed of light and this is how nature enforces the speed limit c.

# Beta (β) Decay

The electron recoils we just mentioned occur in a process called  $\beta$  (beta) decay. In a  $\beta$  decay, a radioactive or unstable nucleus transforms into the nucleus of another element by ejecting an electron at high speeds as illustrated in Figure (7). In the process the nucleus itself recoils as shown.



## Figure 7

Radioactive decay of a nucleus by  $\beta$  decay. In this process the unstable nucleus ejects an electron, often at speeds  $v_e$  near the speed of light. The name  $\beta$  decay is historical in origin. When Ernest Rutherford (who later discovered the atomic nucleus) was studying radioactivity in the late 1890s, he noticed that radioactive materials emitted three different kinds of radiation or rays, which he arbitrarily called  $\alpha$  (alpha) rays,  $\beta$  (beta) rays and  $\gamma$  (gamma) rays, after the first three letters of the Greek alphabet. Further investigation over the years revealed that  $\alpha$  rays were beams of helium nuclei, which are also known as  $\alpha$  particles. The  $\beta$  rays turned out to be beams of electrons, and for this reason a nuclear decay in which an electron is emitted is known as a **\beta decay**. The  $\gamma$  rays turned out to be particles of light which we now call photons. (The particle nature of light will be discussed in a later section of this chapter.)

In the 1920s, studies of the  $\beta$  decay process raised serious questions about some fundamental laws of physics. It appeared that in the  $\beta$  decay, energy was sometimes lost. (We will discuss energy and the basic law of conservation of energy in Chapter 9.) In the early 1930s, Wolfgang Pauli proposed that in  $\beta$  decay, two particles were emitted—an electron and an undetectable one which later became known as the neutrino. (We will discuss neutrinos at the end of this chapter.) Pauli's hypothesis was that the missing energy was carried out by the unobservable neutrino. Thirty years later the neutrino was finally detected and Pauli's hypothesis verified.

Some of the time the neutrino created in a  $\beta$  decay carries essentially no energy and has no effect on the behavior of the electron and the nucleus. When this is the case, we have the genuine 2-particle recoil experiment illustrated in Figure (7). This is a recoil experiment in which one of the particles emerges at speeds near the speed of light.

## Electron Mass in $\beta$ Decay

Applying our definition of mass to the  $\beta$  decay process of Figure (7) we have



where  $m_e$  and  $v_e$  are the mass and recoil speed of the electron and  $m_n$  and  $v_n$  of the nucleus. We are assuming that the nucleus was originally at rest before the  $\beta$  decay.

To develop a feeling for the speeds and masses involved in the  $\beta$  decay process, we will analyze two examples of the  $\beta$  decay of a radioactive nucleus. In the first example, which we introduce as an exercise to give you some practice calculating with Equation (5), we can assume that the electron's mass is unchanged and still predict a reasonable speed for the ejected electron. In the second example, the assumption that the electron's mass is unchanged leads to nonsense.

# **Plutonium 246**

We will begin with the decay of a radioactive nucleus called *Plutonium 246*. This is not a very important nucleus. We have selected it because of the way in which it  $\beta$  decays.

The number 246 appearing in the name tells us the number of protons and neutrons in the nucleus. Protons and neutrons have approximately the same mass  $m_p$  which has the value

$$m_{\rm p} = 1.67 \times 10^{-27} \, \text{kg} \quad \text{mass of proton} \tag{6}$$

The Plutonium 246 nucleus has a mass 246 times as great, thus

$$m_{Plutonium 246} = 246 \times m_p$$
  
= 4.10 × 10<sup>-25</sup> kg (7)

An electron at rest or moving at slow speeds has a mass  $(m_e)_0$  given by

$$(m_e)_0 = 9.11 \times 10^{-31} \,\mathrm{kg}$$
 (8)

This is called the *rest mass* of an electron. We have added the subscript zero to remind us that this is the mass of a slowly moving electron, one traveling at speeds much less than the speed of light.

# Exercise 2 $\beta$ Decay of Plutonium 246

A Plutonium 246 nucleus has an average lifetime of just over 11 days, upon which it decays by emitting an electron. If the nucleus is initially at rest, and the decay is one in which the neutrino plays no role, then the nucleus will recoil at the speed

$$v_n = 572 \frac{\text{meters}}{\text{second}}$$
  $recoil speed of$   
Plutonium 246 (9)  
 $in a \beta decay$ 

This recoil speed is not observed directly, but enough is known about the Plutonium 246  $\beta$  decay that this number can be accurately calculated. Note that a speed of 572 meters/second is a bit over 1000 miles per hour, the speed of a supersonic jet.

Your exercise is to predict the recoil speed  $v_e$  of the electron assuming that the mass of the electron  $m_e$  is the same as the mass  $(m_e)_0$  of an electron at rest.

Your answer should be

$$v_e = .86 c$$
 (10)

where

$$c = 3 \times 10^8 \frac{\text{meters}}{\text{sec ond}}$$
(11)  
is the speed of light.

The above exercise, which you should have done by now, shows that we do not get into serious trouble if we assume that the mass of the electron did not change due to the electron's motion. The predicted recoil speed  $v_e = .86c$  is a bit too close to the speed of light for comfort, but the calculation does not exhibit any obvious problems. This is not true for the following example.

# **Protactinium 236**

An even more obscure nucleus is Protactinium 236 which has a lifetime of about 12 minutes before it  $\beta$  decays. The Protactinium  $\beta$  decay is, however, much more violent than the Plutonium 246 decay we just discussed. If the Protactinium 236 nucleus is initially at rest, and the neutrino plays no significant role in the decay, then the recoil velocity of the nucleus is

 $v_n = 5170 \frac{\text{meters}}{\text{second}}$   $\frac{recoil speed of}{Protactinium 236}$  (12)

This is nine times faster than the recoil speed of the Plutonium 246 nucleus.

# Exercise 3 Protactinium 236 $\beta$ decay.

Calculate the recoil speed of the electron assuming that the mass of the recoiling electron is the same as the mass of an electron at rest. What is wrong with the answer?

You do not have to work Exercise 3 in detail to see that we get a into trouble if we assume that the mass of the recoiling electron is the same as the mass of an electron at rest. We made this assumption in Exercise 2, and predicted that the electron in the Plutonium 246  $\beta$ decay emerged at a speed of .86 c. Now a nucleus of about the same mass recoils 9 times faster. If the electron mass is unchanged, it must also recoil 9 times faster, or over seven times the speed of light. This simply does not happen.

#### Exercise 4 Increase in Electron Mass.

Reconsider the Protactinium 236 decay, but this time assume that the electron emerges at essentially the speed of light ( $v_e = c$ ). (This is not a bad approximation, it actually emerges at a speed v = .99c). Use the definition of mass, Equation 5, to calculate the mass of the recoiling electron. Your answer should be

$$m_e = 6.8 \times 10^{-30} kg = 7.47 \times (m_e)_0$$
 (13)

In Exercise 4, you found that by assuming the electron could not travel faster than the speed of light, the electron mass had increased by a factor of 7.47. The emerging electron is over 7 times as massive as an electron at rest! Instead of emerging at 7 times the speed of light, the electron comes out with 7 times as much mass.

## Exercise 5 A Thought Experiment.

To illustrate that there is almost no limit to how much the mass of an object can increase, imagine that we perform an experiment where the earth ejects an electron and the earth recoils at a speed of 10 cm/sec. (A  $\beta$  decay of the earth.) Calculate the mass of the emitted electron. By what factor has the electron's mass increased?

# THE EINSTEIN MASS FORMULA

A combination of the recoil definition of mass with the observation that nothing can travel faster than the speed of light, leads to the conclusion that the mass of an object must increase as the speed of the object approaches the speed of light. Determining the formula for how mass increases is a more difficult job. It turns out that we do not have enough information at this point in our discussion to derive the mass formula. What we have to add is a new basic law of physics called the *law of conservation of linear momentum*.

We will discuss the conservation of linear momentum in the next chapter, and in the appendix to that chapter, derive the formula for the increase in mass with velocity. We put the derivation in an appendix because it is somewhat involved. But the answer is very simple, almost what you might guess.

In our discussion of moving clocks in Chapter 1, we saw that the length T' of the astronaut's second increased according to the formula

$$T' = \frac{T}{\sqrt{1 - v^2/c^2}}$$
(1-11)

where T was the length of one of our seconds. For slowly moving astronauts where v << c, we have T'  $\approx$  T and the length of the astronaut's seconds is nearly the same as ours. But as the astronaut approaches the speed of light, the number  $\sqrt{1-v^2/c^2}$ becomes smaller and smaller, and the astronaut's seconds become longer and longer. If the astronaut goes at the speed of light,  $1/\sqrt{1-v^2/c^2}$  becomes infinitely large, the astronaut's seconds become infinitely long, and time stops for the astronaut. Essentially the same formula applies to the mass of a moving object. If an object has a mass  $m_0$  when at rest or moving slowly as in air cart experiments (we call  $m_0$  the *rest mass* of the object), then when the object is moving at a speed v, its mass m is given by the formula

$$m = \frac{m_o}{\sqrt{1 - v^2/c^2}} \qquad \begin{array}{c} Einstein \\ mass \\ formula \end{array}$$
(14)

a result first deduced by Einstein.

Equation (14) has just the properties we want. When the particle is moving slowly as in our air cart recoil experiments,  $v \ll c$ ,  $\sqrt{1-v^2/c^2} \approx 1$  and the mass of the object does not change with speed. But as the speed of the object approaches the speed of light, the  $\sqrt{1-v^2/c^2}$ approaches zero, and m =  $m_0/\sqrt{1-v^2/c^2}$  increases without bounds. If we could accelerate an object up to the speed of light, it would acquire an infinite mass.

## **Exercise 6**

At what speed does the mass of an object double (i.e., at what speed does  $m = 2 m_0$ ?) (Answer: v = .866 c.)

## **Exercise 7**

Electrons emerging from the Stanford Linear Accelerator have a mass 200,000 times greater than their rest mass. What is the speed of these electrons? (The answer is v = .9999999999875 c. Use the approximation formulas discussed in Chapter 1 to work this problem.)

#### **Exercise 8**

A car is traveling at a speed of v = 68 miles per hour. (68 miles/hr = 100 ft/second =  $10^{-7}$  ft/nanosecond =  $10^{-7}$  c.) By what factor has its mass increased due to its motion. (Answer:  $m/m_o = 1.0000000000005$ .)

# **Nature's Speed Limit**

When the police try to enforce a 65 mile/hr speed limit, they have a hard job. They have to send out patrol cars to observe the traffic, and chase after speeders. Even with the most careful surveillance, many drivers get away with speeding.

Nature is more clever in enforcing its speed limit c. By having the mass of an object increase as the speed of the object approaches c, it becomes harder and harder to change the speed of the object. If you accelerated an object up to the speed of light, its mass would become infinite, and it would be impossible to increase the particle's speed.

Historically it was noted that massive objects were hard to get moving, but when you got them moving, they were hard to stop. This tendency of a massive object to keep moving at constant velocity was given the name *inertia*. That is why our recoil definition of mass, which directly measures how hard it is to get an object moving, measures what is called *inertial mass*. Nature enforces its speed limit c by increasing a particle's inertia to infinity at c, making it impossible to accelerate the particle to higher speeds. Because of this scheme, no one speeds and no police are necessary.

# ZERO REST MASS PARTICLES

If you think about it for a while, you may worry that nature's enforcement of its speed limit c is too effective. With the formula  $m = m_0/\sqrt{1-v^2/c^2}$ , we expect that nothing can reach the speed of light, because it would have an infinite mass, which is impossible.

What is light? It travels at the speed of light. If light consists of a beam of particles, and these particles travel at the speed c, then the formula  $m = m_0/\sqrt{1-v^2/c^2}$  suggests that these particles have an infinite mass, which is impossible.

Then perhaps light does not consist of particles, and is therefore exempt from Einstein's formula. Back in Newton's time there was considerable debate over the nature of light. Isaac Newton supported the idea that light consisted of beams of particles. Red light was made up of red particles, green light of green particles, blue light of blue particles, etc. Christian Huygens, a well known Dutch physicist of the time, proposed that light was made up of waves, and that the different colors of light were simply waves with different wavelengths. Huygens developed the theory of wave motion in order to support his point of view. We will discuss Huygen's theory later in the text.

In 1801, about 100 years after the time of Newton and Huygens, Thomas Young performed an experiment that settled the debate they started. With his so called *two slit experiment* Young conclusively demonstrated that light was a wave phenomena.

Another century later in 1905, the same year that he published the special theory of relativity, Einstein also published a paper that conclusively demonstrated that light consisted of beams of particles, particles that we now call *photons*. (Einstein received the Nobel Prize in 1921 for his paper on the nature of light. At that time his special theory of relativity was still too controversial to be awarded the prize.)

Thus by 1905 it was known that light was both a particle and a wave. How this could happen, how to picture something as both a particle and a wave was not understood until the development of quantum mechanics in the period 1923 through 1925.

Despite the fact that light has a wave nature, it is still made up of beams of particles called photons, and these particles travel at precisely the speed c. If we apply Einstein's mass formula to photons, we get for the photon mass  $m_{photon}$ 

$$m_{\text{photon}} = \left. \frac{m_0}{\sqrt{1 - v^2/c^2}} \right|_{v=c} = \frac{m_0}{\sqrt{1 - 1}} = \frac{m_0}{0}$$
(15)

where  $m_0$  is the rest mass of the photon.

At first sight it looks like we are in deep trouble with Equation (15). Division by zero usually leads to a disaster called infinity. There is one exception to this disaster. If the rest mass  $m_0$  of the photon is zero, then we get

$$m_{\text{photon}} = \frac{m_0}{0} = \frac{0}{0} \tag{16}$$

The number 0/0 is not a disaster, it is simply undefined. It can be 1 or 2.7, or  $6 \times 10^{-23}$ . It can be any number you want. (How many nothings fit into nothing? As many as you want.) In other words, if the rest mass  $m_0$  of a photon is zero, the Einstein mass formula says nothing about the photon's mass  $m_{photon}$ . Photons do have mass, but the Einstein mass formula does not tell us what it is. (Einstein presented a new formula for the photon's mass was proportional to the frequency of the light wave.)

We will study Einstein's theory of photons in detail later in the text. All we need to know now is that light consists of particles called photons, these particles travel <u>at</u> the speed of light, and these particles have no rest mass. If you stop photons, which you do all the time when light strikes your skin, no particles are left. There is no residue of stopped photons on your skin. All that is left is the heat energy brought in by the light.

A photon is an amazing particle in that it exists only when moving at the speed of light. There is no lapse of time for photons; they cannot become old. (They cannot spontaneously decay like muons, because their half life would be infinite.) There are two different worlds for particles. Particles with rest mass cannot get up to the speed of light, while particles without rest mass travel only at the speed of light.

# **NEUTRINOS**

Another particle that may have no rest mass is the neutrino. According to current theory there should be three different kinds of neutrinos, but for now we will not distinguish among them.

In our discussion of the  $\beta$  decay process, we mentioned that when a radioactive nucleus decays by emitting an electron, a neutrino is also emitted. Most of the time the energy given up by the nucleus is shared between the electron and the neutrino, thus the electrons carried out only part of the energy. The very existence of the neutrino was predicted from the fact that some energy appeared to be missing in  $\beta$  decay reactions and it was Pauli who suggested that this energy was carried out by an undetected particle.

Neutrinos are difficult to detect. They can pass through immense amounts of matter without being stopped or deflected. In comparison photons are readily absorbed by matter. As any scuba diver knows, even in the clearest ocean, a good fraction of the sunlight is absorbed by the time you get down to a depth of 50 or more feet. At that depth most of the red light has been absorbed and objects have a grayish blue cast. In muddy water photons are absorbed much more rapidly, and opaque objects like your skin stop photons in the distance of a few atomic diameters.

On the other hand, neutrinos can pass through the earth with almost no chance of being stopped. As a writer discussing the 1987 supernova explosion phrased it, the neutrinos from the supernova explosion swept through the earth, the earth being far more transparent to the neutrinos than a thin sheet of the clearest glass to light. Neutrinos are now detected by what one might call a brute force technique. Aim enough neutrinos at a big enough detector and a few will be stopped and observed. The first time neutrinos were detected was in an experiment by Clyde Cowan and Fred Reines, performed in 1956, almost 30 years after Pauli had proposed the existence of the particle. Noting that nuclear reactors are a prodigious source of neutrinos, Cowan and Reines succeeded in detecting neutrinos by building a detector the size of a railroad tank car and placing it next to the reactor at Savannah River, Georgia.

The largest neutrino detectors now in use were originally built to detect the spontaneous decay of the proton (a process that has not yet been observed). They consist of a swimming pool sized tank of water surrounded by arrays of photocells, all located in deep mines to shield them from cosmic rays. If a proton decays, either spontaneously or because it was struck by a neutrino, a tiny flash of light is emitted in the subsequent particle reaction. The flash of light is then detected by one of the photocells.

## **Solar Neutrinos**

Aside from nuclear reactors, another powerful source of neutrinos is the sun. Beta decay processes and neutrino emission are intimately associated with the nuclear reactions that power the sun. As a result neutrinos emerge from the small hot core at the center of the sun where the nuclear reactions are taking place. The sun produces so many neutrinos that we can detect them here on earth. There is a good reason to look for these solar neutrinos. The neutrinos created in the core of the sun pass directly through the outer layers of the sun and reach us eight minutes after they were created in a nuclear reaction. In contrast, light from the hot bright core of the sun takes the order of 14,000 years to diffuse its way out to the surface of the sun. If for some reason the nuclear reactions in the sun slowed down and the core cooled, it would be about 14,000 years before the surface of the sun cooled. But the decrease in neutrinos could be detected here on earth within 8 minutes. Looking at the solar neutrinos provides a way of looking at the future of the sun 14,000 years from now.

Solar neutrinos have been studied and counted since the 1960s. Computer models of the nuclear reactions taking place in the sun make explicit predictions about how many neutrinos should be emitted. The neutrino detectors observe only about 1/3 to 1/2 that number. There have been a number of experiments using various kinds of detectors, and all the experiments show this deficiency.

If the deficiency is really an indication that the nuclear reactions in the sun's core have slowed, then we can expect a cooling of the sun within 14,000 years, a cooling that might have a significant impact on the earth's climate. On the other hand there may be some part of the nuclear reactions in the sun that we do not fully understand, with the result that the computer predictions are in error. We are not sure yet which is correct; the solar neutrino deficiency is one of the current areas of active research.

# **Neutrino Astronomy**

An event on the night of February 23, 1987 changed the role of neutrinos in modern science. On that night neutrinos were detected from the supernova explosion in the Magellanic cloud, a small neighboring galaxy. This was the first time neutrinos were detected from an astronomical source other than our sun. The information we obtained from this observation represented what one could call the birth of neutrino astronomy.

A supernova is an exploding star, an event so powerful that, for a short period of time of about 10 seconds, the star radiates more power than all the rest of the visible universe. And this energy is radiated in the form of neutrinos.

The supernova explosion occurs when the core of a large star runs out of nuclear fuel and collapses. (This only happens to stars several times larger than our sun.) The gravitational energy released in the collapse is what provides the energy for the explosion. We know that sometimes a neutron star is formed at the center of the collapsed core, and computer simulations predict



**Figure 8** 1987 Supernova at age  $3\frac{1}{2}$  years, photographed by the Hubble telescope. The ring is gas blown off by the explosion.

that much of the energy released in the collapse is carried out in a burst of neutrinos. The core material is so dense that even the neutrinos have some difficulty getting out. They take about 10 seconds to diffuse out of the core, and as a result the neutrino pulse is about 10 seconds long.

The collapsing core also creates a shock wave that spreads out through the outer layers of the star, reaching the surface in about three hours. When the shock wave reaches the surface, the star suddenly brightens and we can see from the light that the star has exploded.

The details about the core collapse, the neutrino burst and the shock wave are all from computer models of supernova explosions, models developed over the past 25 years. Whenever you model a physical process, you like to test your model with the real process. Computer models of supernova explosions are difficult to test because there are so few supernova explosions. The last explosion in our galaxy, close enough to study in detail, occurred in 1604, shortly before the invention of the telescope.

The supernova explosion on February 23, 1987 was not only close enough to be studied, several fortunate coincidences provided much detailed information. The first coincidence was the fact that theoretical physicists had predicted in the 1960s that the proton might spontaneously decay (with a half life of about  $10^{32}$ years.) To detect this weak spontaneous decay, several large detectors were constructed. As we mentioned, these large detectors were also capable of detecting neutrinos. On February 23, at 7:36 AM universal time, the detectors in the Kamokande lead mine in Japan, the Morton Thekol salt mine near Cleveland, Ohio and at Baksam in the Soviet Union all detected a 10 second wide pulse of neutrinos. Since the Magellanic cloud and the supernova are visible only from the southern hemisphere and all the neutrino detectors are in the northern hemisphere, all the detected neutrinos had to pass through the earth. The 10 second width of the pulse verified earlier computer models about the diffusion of neutrinos out of the collapsing core.

The exact time of the arrival of the light from the supernova explosion is harder to pin down, but some fortunate coincidences occurred there too. The supernova was first observed by a graduate student Ian Sheldon working at the Las Campanas Observatory in Chile. Ian was photographing the large Magellanic cloud on the night of February 23, 1987, and noted that a plate that he had exposed that night had a bright stellar object that was not on the plate exposed the night before. The object was so bright it should be visible to the naked eye. Ian went outside, looked up, and there it was.

Once the supernova had been spotted, there was an immediate search for more precise evidence of when the explosion had occurred. A study of the records of the neutrino detectors turned up the ten second neutrino pulse at 7:36 AM on February 23. Three hours after that Robert McNaught, an observer in Siding Spring, Australia, had exposed two plates of the large Magellanic clouds. When the plates were developed later, the supernova was visible. One hour before McNaught exposed his plates, Albert Jones, an amateur astronomer in New Zealand happened to be observing at the precise spot where the supernova occurred and saw nothing unusual. Thus the light from the supernova explosion arrived at some time between two and three hours after the neutrino pulse.

The fact that the photons from the supernova explosion arrived two to three hours later than the neutrinos, is not only a good test of the computer models of the supernova explosion, it also provides an excellent check on the rest mass of the neutrino. The 1987 supernova occurred 160,000 light years away from the earth. After the explosion, neutrinos and photons raced toward the earth. The neutrinos had a 3 hour head start, and after traveling for 160,000 years, the neutrinos were still 2 hours, and perhaps 3 hours ahead. That is as close a race as you can expect to find. From this we can conclude that neutrinos travel at, or very, very close to the speed of light. And therefore their mass must be precisely zero, or very close to it.

# Chapter 7 **Conservation of Linear and Angular Momentum**

The truly basic laws of physics, like the principle of relativity, not only have broad applications, but are often easy to describe. The principle of relativity says that there is a quantity, namely your own uniform motion, that you cannot detect. The hard part is working out the implications of the simple idea.

In this chapter we discuss two more basic laws of physics, laws that apply with no known exceptions to objects as large as galaxies and as small as subatomic particles. These are the **laws of the conser**vation of linear momentum and the **law of the** conservation of angular momentum. These are the first of several so-called conservation laws that we will encounter in our study of physics. A conservation law states that there is some quantity which does not change in a given set of experiments.

We will introduce our first example of a conservation law by going back to the results of the aircart recoil experiments that we used in the last chapter to define mass. In analyzing these results, we will see that there is a quantity, which we will call linear momentum, which does not change when the spring is released and the carts recoil. We will then look at a wider class of experiments, in which objects not only recoil, but collide at different angles. Again we will see that linear momentum does not change.

We will also see that linear momentum is conserved not just for familiar objects like billiard balls, but also for objects as small as protons colliding in a hydrogen bubble chamber.

In the appendix to this chapter we will show how the recoil definition of mass, when combined with the law of conservation of linear momentum and the principle of relativity, leads to Einstein's relativistic mass formula  $m = m_0 / \sqrt{1 - v^2/c^2}$ .

The second conservation law deals with angular momentum. The concept of angular momentum is a bit more subtle than that of linear momentum. As a result, in this chapter we will focus on developing an intuitive feeling for the concept. A more formal mathematical treatment will be put off until a later chapter where the formalism is needed. The law of conservation of angular momentum has many applications that range from the astronomical scale to the subatomic scale of distance. The very existence of planets in the solar system is a consequence of the conservation of angular momentum. The law also allows us to understand the behavior of atomic nuclei in a magnetic field, a behavior that is involved in the creation of the marvelous images seen in magnetic resonance imaging apparatus. On the very smallest scale of distance, angular momentum turns out to be one of the basic intrinsic properties of all elementary particles.

# CONSERVATION OF LINEAR MOMENTUM

In our discussion of the recoil definition of mass in the last chapter, we looked at a number of experiments that were performed to determine how mass behaves. One of the crucial observations was that, at least with carts on an air track, the ratio of the recoil speeds did not change as we changed the strength of the spring pushing the carts apart. If the big cart came out moving half as fast as the small one when we used a weak spring, then it still came out half as fast when we used a strong spring. With the strong spring both speeds were greater, but the ratio was still the same.

We used this unchanging ratio in our definition of mass. If, as shown in Figure (6-3) reproduced here, cart A recoils at a speed  $v_A$  and cart B at a speed  $v_B$ , then the mass ratio  $m_A/m_B$  was defined by Equation (6-1) as

$$\frac{\mathbf{m}_{\mathrm{A}}}{\mathbf{m}_{\mathrm{B}}} = \frac{\mathbf{v}_{\mathrm{B}}}{\mathbf{v}_{\mathrm{A}}} \tag{6-1}$$

What we will do now is manipulate Equation (6-1) until we end up with a quantity that does not change when the spring is released.





In Chapter 6 we defined the ratio of the mass of the carts  $m_A/m_B$  to be equal to the inverse ratio of the recoil speeds  $v_B/v_A$ .

Multiplying Equation (6-1) through by  $m_B$  and  $v_A$  gives

$$m_A v_A = m_B v_B \tag{1}$$

Next, note that  $v_A$  and  $v_B$  are the magnitudes of the recoil velocity vectors  $\vec{v}_A$  and  $\vec{v}_B$ . Since  $\vec{v}_A$  and  $\vec{v}_B$  point in opposite directions, we can write Equation (1) as a vector equation in the form

$$m_A \vec{v}_A = -m_B \vec{v}_B \tag{2}$$

where the minus sign handles the fact that  $\vec{v}_A$  and  $\vec{v}_B$  are oppositely directed.

Now move the  $-m_B \vec{v}_B$  to the left hand side to give the result

$$m_{\rm A}\vec{\rm v}_{\rm A} + m_{\rm B}\vec{\rm v}_{\rm B} = 0 \quad after \ recoil \tag{3}$$

where  $\vec{v}_A$  and  $\vec{v}_B$  are the cart's velocity vectors after the recoil.

We will now introduce a new interpretation. Let us define the *linear momentum* of a particle as the product of the particle's mass times its velocity. Using the letter p to denote linear momentum, we have

$$\vec{p} \equiv m\vec{v} \quad \begin{array}{c} definition of \\ linear momentum \end{array}$$
(4)

Note that linear momentum  $\vec{p}$  is a vector because it is the product of a number, the mass m, times a vector, the velocity  $\vec{v}$ .

Looking back at Equation (3), we see that  $m_A \vec{v}_A$  is the linear momentum of cart A after the recoil, and  $m_B \vec{v}_B$  is the linear momentum of cart B after the recoil. Equation (3) tells us that the sum of these two linear momenta is zero.

Before the string was cut and the carts released, both carts were sitting at rest. Before the release, we have

$$\vec{v}_A = 0$$
,  $\vec{v}_B = 0$  before release (5)

Thus the sum of the linear momenta before the release was

$$m_{A}\vec{v}_{A} + m_{B}\vec{v}_{B} = 0 \quad before \ recoil \tag{6}$$

Comparing Equations (3) and (6) we see that the sum of the linear momenta of the two carts was not changed by the recoil. This sum was zero before the carts were released, and it is zero afterward.

We will call the sum of the linear momentum of the two carts the total linear momentum  $\vec{p}_{tot}$  of the system of two carts.

$$\vec{p}_{tot} \equiv \vec{p}_A + \vec{p}_B = m_A \vec{v}_A + m_B \vec{v}_B$$
*definition of total linear momentum*
(7)

Then we can restate our observation that Equation (3) and (6) look the same by saying that the total linear momentum  $\vec{p}_{tot}$  of the system was unchanged by the recoil. Another way of phrasing it is to say that in the recoil experiment, the total linear momentum of the carts is conserved.

At this point, we do not have a new law of physics, instead, we have merely reformulated our definition of mass. But the result turns out to be far more general than we have seen so far. The general law may be stated as follows. *If we have a system of particles, and there is no net external force acting on them, then the total linear momentum of the system of particles is conserved.* 

So far, we have not said much about forces and how to recognize them. Thus we will, for now, limit our discussion of the conservation of linear momentum to examples where it is fairly clear that there is no net external force or influence. In our recoil experiment, gravity is pulling down on the carts, the air is pushing up, and the two effects cancel. The air track was explicitly designed so that there would be no net force on the cart.

In our recoil experiment, our system consists of the two carts and the spring. When the thread is burned, the spring exerts a force on both carts, but the spring is part of the system. The spring forces are internal, not external forces, and therefore cannot change the linear momentum of the system. If the linear momentum was zero before the thread was burned, it must still be zero afterward.



Figure 1

Experimental setup to study the conservation of linear momentum during the collision of two balls.

# **COLLISION EXPERIMENTS**

A more common example of where external forces can be ignored and linear momentum is conserved is during the collision of two objects like billiard balls. While two objects are colliding, the forces between the objects, the internal forces, are usually much greater than any outside external forces. As a result, just before, during, and just after the collision, external forces can be neglected and linear momentum is conserved.

In an experiment, that is easily carried out in the introductory physics lab, two steel balls are suspended by strings from the ceiling as shown in Figure (1). One of the two balls is pulled back and released. It strikes the ball at rest and the two balls bounce off as seen in the strobe photograph of the motion, Figure (2). The strobe photograph is analyzed in Figure (3) and the resulting momentum vectors are plotted in Figure (4).

What we see in the strobe photograph is ball 2 at rest and ball 1 coming in, attaining a velocity  $\vec{v}_{1i}$  just before the collision. After the collision, balls 1 and 2 bounce off in different directions, with velocity vectors  $\vec{v}_{1f}$  and  $\vec{v}_{2f}$  respectively.



## Figure 2

Strobe photograph taken using the setup of Figure (1). The data for the experiment are  $m_1 = 70.3 \text{ gm}$  (the ball initially released),  $m_2 = 240 \text{ gm}$  (ball initially at rest),  $\Delta t = 1/10 \text{ sec}$  (period between flashes). Spacing between grid lines = 1 cm. (Photograph from a student lab notebook.)

In Figure (4) we have plotted the momentum  $\vec{p}_{1i}$  of ball 1 just before the collision

$$\vec{p}_{1i} = \vec{mv}_{1i} = ball \ l \ before the collision$$
(8)

and also plotted the momenta  $\vec{p}_{1f}$  and  $\vec{p}_{2f}$  of balls 1 and 2 after the collision. The vector sum of these two momenta is the total momentum  $\vec{p}_f$  being carried out by the two balls

$$\vec{p}_{f} \equiv \vec{p}_{1f} + \vec{p}_{2f} = \begin{array}{c} momentum \ being \\ carried \ out \ of \\ the \ collision \end{array}$$
(9)

From Figure (4) we see that  $\vec{p}_f$  is equal to  $\vec{p}_{1i}$  the momentum brought into the collision by ball 1. Since the same amount of momentum came out of the collision as was carried in, the total linear momentum did not change. The total linear momentum of the system of the two balls was conserved during the collision.



#### Figure 3

Analysis of Figure (2). Ball 1 enters with a momentum  $\vec{p}_{1i}$  and collides with Ball 2 which is initially at rest. After the collision, Balls 1 and 2 emerge with momenta  $\vec{p}_{1f}$  and  $\vec{p}_{2f}$  respectively. (Each large square on this graph paper represents a distance of 10 cm.)



#### Figure 4

Here we see that the momentum  $\vec{p}_{1i}$  brought in by Ball 1 is equal to the momentum  $\vec{p}_{1f} + \vec{p}_{2f}$  carried out after the collision.

# Exercise 1

Figures (5 a and b) show the collision between two balls of equal mass  $m_1 = m_2 = 73 \text{ grams}$ . Again  $\Delta t = 1/10 \text{ sec}$ . Using one of these figures, construct a graph similar to Figure (4), and compare the momentum brought in by Ball 1 with the momentum carried out by the two balls after the collision.



**Figures 5 a, b** Strobe photographs of the collision of two equal mass balls.

# **Subatomic Collisions**

In the study of subatomic particles, you cannot photograph or image the particles themselves, the best you can do is study the tracks left behind in a particle detector. A common particle detector, developed by Don Glaser in the early 1950s, is the bubble chamber. We will discuss the bubble chamber in more detail in later chapters. However the basic idea is that the bubble chamber is filled with liquid hydrogen, and that a charged particle moving through the liquid hydrogen leaves a track that can be made visible as a string of bubbles.

Bubble chambers are used primarily to study the collisions between subatomic particles. In Figure (6) we have a bubble chamber photograph in which an incoming proton from a particle accelerator moves through the liquid hydrogen until it strikes a hydrogen nucleus, namely another proton. The two protons emerge from the collision, coming out at right angles as shown.

After we discuss the law of conservation of energy, we will show that if two identical particles collide, one of them being initially at rest, then if both energy and linear momentum are conserved during the collision the particles must emerge at right angles. Thus we can use the right angle between the emerging proton tracks in Figure (6) as experimental evidence that linear momentum is conserved even among the interactions of subatomic particles.



#### Figure 6

Collision between two protons. When a charged elementary particle passes through the liquid hydrogen in a bubble chamber, it leaves a trail of bubbles that can be photographed. Here we see a proton coming into the picture from the upper left, and striking the nucleus of one of the hydrogen atoms in the liquid hydrogen. The hydrogen nucleus is itself a proton, and after the collision the two protons emerge as shown in the sketch.

The following examples and exercises are chosen to show some of the more practical applications of the conservation of linear momentum.

## **Example 1** Rifle and Bullet

A 2-kilogram rifle fires a 10-gram bullet at a speed of 400 meter/sec. What is the recoil velocity of the gun? In this case, the rifle and bullet are initially at rest and have zero total linear momentum. Just after the bullet leaves the gun, before any external forces have had time to act on the system, the total momentum of the system (gun plus bullet) is still zero. We get

$$\vec{p}_{gun} + \vec{p}_{bullet} = 0$$
  
 $m_g \vec{v}_g = -m_b \vec{v}_b$ 

where the minus sign indicates that  $\vec{v}_g$  is in the opposite direction to the motion of the bullet. Solving for the magnitude  $v_g$  of the recoil velocity, we get

$$v_{g} = \frac{m_{b}}{m_{g}} v_{b} = \frac{10 \text{ gm}}{2000 \text{ gm}} \times 400 \frac{\text{meters}}{\text{sec}}$$
$$v_{g} = 2 \frac{\text{meters}}{\text{sec}}$$

Thus, we see that the initial recoil velocity of the gun is 2 meters/sec.



In this example we applied the law of conservation of linear momentum over such a short time that outside forces did not have time to act on the system. The conservation of linear momentum applies over longer times, but we must enlarge our concept of the system, as seen in Example (2),

# Example 2

A 78 kilogram hunter standing on nearly frictionless ice fires the gun of the preceding example. What is the recoil velocity of the hunter?

Our system now consists of the bullet, gun and hunter. Initially the total linear momentum of the system is zero. After the bullet is fired, and after the gun is firmly lodged against the shoulder of the hunter, the gun and hunter together recoil at a velocity  $v_h$ . Applying the law of conservation of linear momentum, we have (remembering that  $v_g$  now equals  $v_h$ )

$$\vec{p}_{hunter} + \vec{p}_{gun} + \vec{p}_{bullet} = 0$$

$$m_h \vec{v}_h + m_g \vec{v}_h = -m_b \vec{v}_b$$

$$v_h = \frac{m_b}{m_h + m_g} v_b = \frac{10 \text{ gm}}{(78 + 2) \text{ kg}} \times 400 \frac{\text{meters}}{\text{sec}}$$

$$v_h = \frac{10 \text{ gm}}{80,000 \text{ gm}} \times 400 \frac{\text{meters}}{\text{sec}} = \frac{1}{20} \frac{\text{meter}}{\text{sec}}$$

$$v_h = 5 \frac{\text{cm}}{\text{sec}}$$

#### Exercise 2

a) Starting from the preceding two examples, further enlarge the system. Assume that the hunter is standing firmly on the earth when the gun is fired. Taking the point of view that the earth is initially at rest, calculate the recoil velocity of the gun, hunter, and earth. ( $m_{earth} = 6 \times 10^{27}$  gm)

b) After the bullet strikes the ground, what is the velocity of the earth, assuming it was at rest before the gun was fired?

#### Exercise 3 Frictionless Ice

Suppose you are sitting in the middle of a completely frictionless surface, such as an idealized pond of ice. Propose a method of getting out of such a predicament. (Problem from J. Orear, *Fundamental Physics*, Wiley, New York, 1961.)

#### Exercise 4 Bullet and Block

A 10-gram bullet traveling 300 meters/sec strikes and lodges in a 3-kilogram block of wood initially at rest on a pond of ice. What is the final velocity of the block and bullet after the collision?

## Exercise 5 Two Skaters Throwing Ball

Two skaters, each of mass 60 kilogram, are standing a slight distance apart on nearly frictionless ice. Initially at rest, they throw a 1-kilogram ball back and forth between them; each time the ball travels at a speed of 10 meters/sec over the ice.

- (a) What is the recoil velocity of the first skater immediately after he throws the ball for the first time?
- (b) After the second skater catches the ball for the first time, what is his recoil velocity?
- (c) After the second skater has thrown the ball back for the first time, what is his recoil velocity?
- (d) After the ball has made 10 complete round trips and the first skater is holding the ball, what is the velocity of each skater? What is the total momentum of the system of the two skaters and the ball?

#### Exercise 6 Rocket

An 11-ton rocket consists of 10 tons of fuel. If the fuel is discharged as exhaust gasses that travel at an average speed of 1 mile/sec (relative to the earth), how fast will the rocket be traveling when the fuel is used up? Neglect gravity and air resistance. (Hint: Consider the total momentum of all the exhaust gas.)

# CONSERVATION OF ANGULAR MOMENTUM

Anyone who has watched figure skating in the winter Olympics has seen an example of the conservation of angular momentum. When a figure skater like the one shown in Figure (7) starts her spin, her arms are outstretched and she is turning slowly. As she brings her arms in, she turns faster and faster until the maneuver is completed. She starts the spin with a certain amount of angular momentum, and that amount does not change, is conserved, throughout the spin. To understand the concept of angular momentum, we have to see why the skater had the same angular momentum when rotating slowly with her arms outstretched and rotating rapidly with her arms pulled close to her body.

Even those of us who are not skilled figure skaters can repeat the skaters experience of a spin using a rotating platform and two iron dumbbells. When done as a classroom demonstration, this is sometimes known as the "three dumbbell experiment". The instructor stands on the rotating platform and holds the dumbbells out as shown in Figure (8a). A student helps in the demonstration by starting the instructor rotating slowly. The instructor then pulls in his arms and rotates even faster than a figure skater because of the mass in the dumbbells (Figure 8b). However unless the instructor is skilled at this demonstration, he is likely to make a far less graceful exit from the spin than do the Olympic figure skaters .



#### **Figure 7**

Figure skater doing a spin. As the skater pulls her arms in, she turns faster and faster. This is an example of the conservation of angular momentum.



## Figure 8

The "three dumbbell" experiment. The instructor, standing on a platform that is free to rotate, holds two dumbbells out at arm length as shown in (a). With a slight push from a student, the instructor starts to rotate slowly. The instructor then pulls his arms in, and the rotation increases significantly (b). In a more controlled and idealized experiment, we can set a ball swinging in a circle at the end of a string as shown in Figure (9). Let the other end of the string pass down through the small end of a plastic funnel mounted on a board as shown in Figure (9a). If we pull down on the string to reduce the radius r of the circle around which the ball is traveling, the speed v of the ball increases. An analysis of this motion shows a simple result—the product of the radius of the circle times the speed of the ball remains constant. If the ball is initially moving in a circle of radius  $r_1$  and a speed  $v_1$  as shown in Figure (9b), and we reduce the radius to a length  $r_2$  as shown in Figure (9c), the new speed  $v_2$  is given by the equation

$$\mathbf{v}_1 \mathbf{r}_1 = \mathbf{v}_2 \mathbf{r}_2 \tag{10}$$

Since  $r_2$  is smaller than  $r_1$ ,  $v_2$  must be bigger than  $v_1$  to keep the product constant.

The angular momentum of a ball of mass m traveling at a speed v in a circle of radius r is defined to be the product mvr. Using the letter  $\ell$  to represent angular momentum, we have

$$\begin{array}{c} \label{eq:linear} \left[ \ell \equiv mvr \right] & \begin{array}{c} angular \ momentum of \ a \\ mass \ m \ traveling \ at \ a \ speed \\ v \ in \ a \ circle \ of \ radius \ r \end{array} (11) \end{array}$$

In Figure (9a), the ball has an angular momentum

$$\ell_1 = \mathbf{m}\mathbf{v}_1\mathbf{r}_1 \tag{12}$$

after the string is shortened, the angular momentum  $\ell_2$  is

$$\ell_2 = m v_2 r_2 \tag{13}$$

Since the mass m of the ball did not change, and because  $v_1r_1 = v_2r_2$ , we see that

$$\ell_1 = \ell_2 \tag{14}$$

and in this example angular momentum did not change. The ball's angular momentum was conserved while we pulled on the string and the ball sped up. It was also conserved when the figure skater pulled in her arms during the spin, and the instructor pulled in on the dumbbells. The only difference in the three examples is that we have a more complex formula for angular momentum for the figure skater and instructor.

#### **Exercise 7**

What are the dimensions of angular momentum when mass is measured in grams, length in centimeters, and time in seconds?

#### Exercise 8

Figure (9d) is a strobe photograph from a student lab notebook. The string was suddenly pulled down, shortening the radius of the ball's circular orbit. Using this photograph, show that the angular momentum of the ball did not change.

#### **Exercise 9**

Neglecting the mass of the spokes, what is the angular momentum of a bicycle wheel of mass m and radius r, spinning with a period T? (T is the length of time the wheel takes to go around once.)





## Figure 9

A more controlled demonstration of the conservation of angular momentum. One end of a string is tied to a ball of mass m, and the other is fed down through a plastic funnel mounted on the end of a board shown in (a). The ball is then swung in a circle of radius  $r_1$ , and speed  $v_1$  as shown in (b). Then pull down on the free end of the string, to reduce the radius of the circle to  $r_2$ . It takes a fairly strong tug, but the speed of the ball increases to  $v_2$  as shown in (c). From the experimental results shown in (d), you can check that  $r_1v_1 = r_2v_2$ . Since the angular momentum of the ball is proportional to rv, angular momentum was conserved in this experiment. (Photo from lab of G. Sheldon.)
# A MORE GENERAL DEFINITION OF ANGULAR MOMENTUM

The concept of angular momentum applies to more general situations than mass traveling in a circle. For a more general definition of angular momentum, consider the situation shown in Figure (10). In Figure (10a) a ball with linear momentum  $\vec{p} = m\vec{v}$  is traveling along a path that will take it a distance  $r_{\perp}$  from some point labeled O. To make the situation more realistic, imagine that there is a light rod pivoted at point O with a hook at the other end. The length of the rod is  $r_{\perp}$ , so that the hook will just catch the ball as the ball passes by (Figure 10b).

Once the ball has been hooked, it will travel in a circle as shown in Figure (10c). If the rod is perpendicular to the path of the ball when the ball is hooked (as shown in Figure 10a) then there will be no disruption in the speed of the ball and the ball will move around the circle at the same speed v.

Once the ball is traveling in a circle we know that its angular momentum about the pivot O is given by Equation (11) as  $l = mvr_{\perp}$ . In our generalization of the definition of angular momentum, the ball has this same amount of angular momentum before it was hooked as it did after.

The more general definition is as follows. Consider the path of the ball shown in Figure (10a) and let us use the name *lever arm* for the distance of closest approach from the path to the axis at point O. This lever arm is the perpendicular distance to the path, the distance we have labeled  $r_{\perp}$ . Then as our new definition of angular momentum, we say that the magnitude  $\ell$  of the ball's angular momentum is equal to the product of the magnitude of the linear momentum p = mv times the length of the lever arm  $r_{\perp}$ 

$$\ell = \mathrm{pr}_{\perp} \tag{15}$$

Applying Equation (15) to the situation shown in Figure (10a), before the ball is hooked, we see that as the ball heads toward the hook, its initial angular momentum  $l_i$  is

$$\ell_i = p_i r_\perp = m v r_\perp \tag{16}$$

Since this is the same as the angular momentum after the ball is hooked and traveling in a circle, the angular momentum is unchanged, is conserved, during the process of being captured by the hook.



#### Figure 10



When you have a conserved quantity like angular momentum, it takes on a reality that goes beyond the formulas that define it. With our generalized definition of angular momentum, we find that angular momentum can be passed from one object to another. For example suppose a student is standing motionless on the rotating platform as in Figure (11a), and the instructor tosses a softball off to the side of the student as shown. The softball has a lever arm  $r_{\perp}$  and therefore an angular momentum mv  $r_{\perp}$ about the axis of the rotating platform. If the student reaches out and grabs the ball, she acquires the angular momentum of the ball and starts rotating as shown in Figure (11b). If she brings the ball closer in toward her body, she will rotate faster because the ball has a shorter lever arm.



# Figure 11

Student catching angular momentum from a ball thrown off to the side.

#### **Exercise 10**

If you are standing directly over the axis of the platform and a ball is thrown directly toward you, as shown in Figure (12), do you start to rotate after catching the ball? Try the experiment yourself and see if the prediction is correct.





# ANGULAR MOMENTUM AS A VECTOR

Our definition of angular momentum is clearly not yet complete. Even when we are standing on a rotating platform so that we can freely rotate only about the axis of the platform, there are still two different directions we can rotate—clockwise and counter clockwise. The definition of angular momentum must somehow account for these two directions of rotation.

The study of rotations can be complex, particularly if you allow rotations in three dimensions, about any of the three coordinate axes x, y or z. The rotating platform used in Figure (8) and (11) greatly simplifies the situation by restricting our motion to rotation about one axis, the axis of the platform which is conventionally called the z axis as shown in Figure (13).

One way to distinguish positive and negative rotation is by using a right-hand rule. The rule is to point the thumb of your right hand along the positive z axis, and then say that the direction of positive rotation is the direction that the fingers of your right hand curl, as seen in Figure (13). Looking down (with the z axis pointing up toward us), we find that positive rotation is counter clockwise and negative rotation is clockwise.



#### Figure 13

Definitions of the z axis and of positive and negative rotation. For convenience we will call the axis about which our platform rotates the "z axis". To distinguish the two kinds of rotation, we will use the right hand rule. Curl the finger of your right hand in the direction the platform is rotating. If your thumb points up, we say the rotation is positive. If your thumb points down, the rotation is negative.

### Exercise 11

What would be the direction of positive rotation if we used a left hand convention. Draw a sketch and explain.

The next generalization of our definition of angular momentum is best illustrated by the use of the rotating bicycle wheel mounted on a handle as shown in Figure (14). To make an effective demonstration, the tire of the bicycle wheel has been replaced by wire wrapped along the rim of the wheel to give the wheel added mass. If we spin the wheel all the mass m on the rim is moving at the same speed v and has the same lever arm r about the axis of the wheel. Thus the angular momentum of the rotating wheel is  $\ell = mv r$ .

The next step will at first seem arbitrary, and perhaps downright silly. What we are going to do now is to define the angular momentum of the wheel as a vector, of length  $\ell = mv r$ , *pointing along the axis of the wheel*. Which way it points is defined by a right-hand rule. As shown in Figure (14), curl the fingers of your right hand in the direction that the wheel is rotating and the angular momentum vector  $\vec{\ell}$  points along the axis in the direction of your thumb.

This method of turning angular momentum into a vector seems doubly arbitrary. First of all, the angular momentum vector  $\vec{l}$  points perpendicular



#### Figure 14

A further generalization of the concept of angular momentum is to say that it is a vector  $\hat{\ell}$ . The magnitude  $|\hat{\ell}|$  is just our old definition  $mvr_{\perp}$ . We now say that the angular momentum vector  $\hat{\ell}$  points in the direction of the axis of rotation, in the direction given by the right hand rule. (Curl the fingers of your right hand in the direction the wheel is rotating, and your thumb points in the direction of  $\hat{\ell}$ .) to the plane in which the motion of the wheel is occurring, and then one arbitrarily selects a righthanded instead of a left-hand convention to decide which way along the axis the vector  $\vec{\ell}$  should point. It is hard to believe that such arbitrary choices could have any relationship to physical reality.

But it works, as we can easily demonstrate using the bicycle wheel and the rotating platform. In the first demonstration, have a student stand at rest on the rotating platform and let the instructor spin the bicycle wheel and orient it so that the bicycle wheel's angular momentum vector  $\vec{l}$  is pointing up as shown in Figure (15a). The instructor then hands the bicycle wheel, and its angular momentum, to the student as shown.

On the rotating platform motion is restricted to rotation about the z axis (axis of the platform), thus the only component of angular momentum that is of interest is the z component. In Figure (15a) the bicycle has a positive z component of angular momentum  $\ell_z = mvr$ , and the student has none. Thus the total angular momentum of the wheel and student is

$$(\ell_z)_{\text{total}} = (\ell_z)_{\text{bicycle wheel}} + (\ell_z)_{\text{student}}$$
  
= +mvr + 0 (15)  
= +mvr

Figure 15a

A student, at rest on the rotatable platform, is handed a rotating bicycle wheel whose angular momentum vector  $\hat{\ell}$  is up as shown.



The student turns the bicycle wheel over. If angular momentum is conserved, the student must gain an angular momentum  $2\ell$  directed up as shown.



Have the student turn the bicycle wheel upside down as shown in Figure (15b). Now  $\vec{\ell}$  of the bicycle wheel is pointing down so that the bicycle wheel now has a negative z component of angular momentum

$$(\ell_z)_{\text{bicycle wheel}} = -mvr$$
 (16)

Since the total  $\ell_z$  of the student and bicycle wheel must be conserved, the student must gain a positive z component of angular momentum

$$(\ell_z)_{\text{student}} = +2\text{mvr} \tag{17}$$

so that the sum remains + mvr.

What happens when the student turns the bicycle wheel over is that she starts rotating counter clockwise; she gains a positive z component of angular momentum. If at any time she turns the wheel back up, she will stop rotating.



**Figure 15 Movie** *The student turns the bicycle wheel over.* 



#### Figure 16a

The student at rest is handed a bicycle wheel pointed sideways. In this orientation the wheel has no z component of angular momentum. Once the student has the wheel, the z component of the angular momentum of the plus wheel is conserved.



#### Figure 16b

What happens to the student if she turns the wheel up?



Figure 16c What happens to the student if she turns the wheel down?

If the instructor hands the student the rotating wheel oriented horizontally as shown in Figure (16a), and the student is initially at rest, then neither the student or the bicycle wheel have a z component of angular momentum. The total z component is zero and must remain that way no matter what the student does. In the following exercise, you are to predict what will happen if she turns the wheel up or turns it down.

#### Exercise 12

Explain what will happen if the student orients the bicycle wheel up as shown in Figure (16b). What happens when she turns it down as shown in Figure (16c).

#### Exercise 13

The student at rest on the rotating platform is handed a bicycle wheel at rest. She spins the bicycle wheel and orients it so that the bicycle wheel's angular momentum vector points up. Explain carefully what happens to the student.

If you watch, or better yet try these angular momentum demonstrations yourself, you begin to believe that there is really a quantity called angular momentum that you can pass around and manipulate. Now our emphasis is on gaining an intuitive feeling for the concept, later we will come back to the topic with more mathematical machinery. But it is important to already have an intuitive grasp of the concept or the mathematical machinery will not make sense.



**Figure 17** Detail of Hubble photograph of the Eagle nebula. Each nub is a star surrounded by its own gas cloud. (See page 18 for a more complete photograph of the nebula.)

# **Formation of Planets**

Applications of the law of conservation of angular momentum are not confined to classroom demonstrations. We will end this chapter with a discussion of two astronomical applications. One deals with the formation of planets, and the other their motion about the sun.

Stars are formed in the large clouds of gas that stretch throughout the galaxy. An example of such a gas cloud is the Eagle Nebula shown in Figure (17). A particularly active area of star formation is in the nebula in the constellation of Orion shown in Figure (18). The Orion constellation rises after sunset in early winter, and the nebula is located in the middle of the sword dangling from the three bright stars of Orion's belt. Using binoculars, one can see the nebula as a bright patch of gas. The gas is illuminated by the newly formed stars inside.

A star forms when a lump of gas in the cloud begins to collapse due to the gravitational attraction between the gas particles. Judging from the spacing between stars in the neighborhood of the sun, the sun was formed from the collapse of a region of gas about two light years in radius. As the cloud collapses the gravitational attraction between the particles becomes stronger and stronger. The particles rush toward each other at a faster and faster rate, and finally collapse into a hot ball. The more mass in the collapse, the more gravitational energy released, and the hotter the resulting ball. If the ball is of the order of one tenth the mass of the sun or larger, the ball will be hot enough to ignite nuclear reactions and a star is born.

The lumpiness in the gas clouds is probably related to a large scale turbulence in the flow of the gas in the cloud. Such lumpines is usually associated with rotational motion (vorticies), thus as a lump of gas breaks away from the rest of the cloud, it is likely to have some rotation. While the rotational velocity may be small for a two light year diameter lump, as the lump contracts, the speed increasas due to the conservation of angular momentum. If there is enough angular momentum in the gas cloud, a point is reached where the rotation inhibits further collapse of the cloud.

The only way the cloud can continue to collapse is to leave some mass and most of the angular momentum outside. Computer models indicate that a rotating disk of gas forms outside the newly born star, a disk that contains most of the original angular momentum. After a while the gas in the disk condenses into planets that orbit the star. As a result of being formed from a rotating disk of gas, we expect most planets to lie in a plane, and go around the star in the same direction, which the sun and planets do. If the collapsing gas had no angular momentum, the disk would not form and there would be no planets.



**Figure 18** The nebula in the constellation of Orion. This is a particularly active area for the formation of new stars.



**Figure 18a** Hubble photograph of forming stars in the Orion nebula. They are still surrounded by clouds of gas.

#### Exercise 14

Most of the angular momentum of the solar system is taken up by the distant massive planets Jupiter, Saturn and Uranus. If Jupiter were originally formed from a ring of dust 2 light-years in radius, what must have been the initial rotational speed of these particles? (The distance from the sun to Jupiter is 43 light-minutes or  $2.6 \times 10^3$  light-seconds. Jupiter travels at an orbital speed of  $1.3 \times 10^6$  cm/sec for a period of nearly 12 years. 1 year  $\approx \pi \times 10^7$  sec )



#### Figure 17b

Hubble photograph of the Eagle nebula. The nubs at the top of the tallest column are young stars with their own gas clouds. Extremely bright light from a star in the background is pushing away all gas that is not gravitationally attached to a star.

# Chapter 8 Newtonian Mechanics

In Chapters 4 and 5 we saw how to use calculus and the computer in order to predict the motion of a projectile. We saw that if we knew the initial position and velocity of an object, and had a formula for its acceleration vector, then we could predict its position far into the future.

To go beyond a discussion of projectile motion, to develop a general scheme for predicting motion, two new concepts are needed. One is mass, discussed in chapter 6, and the other is force, to be introduced now. We will see that once we know the forces acting on an object, we can obtain a formula for the object's acceleration and then use the techniques of Chapters 4 and 5 to predict motion. This scheme was developed in the late 1600s by Isaac Newton and is known as Newtonian Mechanics.

# FORCE

The concept of a force—a push or a pull—is not as strange or unfamiliar as the acceleration vector we have been discussing. When you push on an object you are exerting a force on that object. The harder you push, the stronger the force. And the direction you push is the direction of the force. From this we see that force is a quantity that has a magnitude and a direction. As a result, it is reasonable to assume that a force is described mathematically by a vector, which we will usually designate by the letter  $\vec{F}$ .

It is often easy to see when forces are acting on an object. What is more subtle is the relationship between force and the resulting acceleration it produces. If I push on a big tree, nothing happens. I can push as hard as I want and the tree does not move. (No bulldozers allowed.) But if I push on a chair, the chair may move. The chair moves if I push sideways but not if I push straight down.

The ancient Greeks, in particular, Aristotle, thought that there was a direct relationship between force and velocity. He thought that the harder you pushed on an object, the faster it went. There is some truth in this if you are talking about pushing a stone along the ground or pulling a boat through water. But these examples, which were familiar problems in ancient time, turn out to be complex situations, involving friction and viscous forces.



The earth's gravitational force produces a uniform downward gravitational acceleration. (Figure 3-27)

Only when Galileo focused on a problem without much friction – projectile motion – did the important role of the acceleration vector become apparent. Later, Newton compared the motion of a projectile (the apple that supposedly fell on his head) with the motion of the planets and the moon, giving him more examples of motion without friction. These examples led Newton to the discovery that force is directly related to acceleration, not velocity.

In our discussion of projectile motion, and projectile motion with air resistance, we have begun to see the relation between force and acceleration. While a projectile is in flight, and we can neglect air resistance, the projectile's acceleration is straight down, in the direction of the earth as shown in Figure (1). As we stand on the earth, we are being pulled down by gravity. While the projectile is in flight, it is also being pulled down by gravity. It is a reasonable guess that the projectile's downward acceleration vector  $\vec{g}$  is caused by the gravitational force of the earth.

When we considered the motion of a particle at constant speed in a circle as shown in Figure (2), we saw that the particle's acceleration vector pointed toward the center of the circle. A simple physical example of this circular motion was demonstrated when we tied a golf ball to a string and swing it over our head.



**Figure 2** The acceleration of the ball is in the same direction as the force exerted by the string. (Figure 3-28)

While swinging the golf ball, it was the string pulling on the ball that kept the ball moving in a circle. (Let go of the string and the ball goes flying off.) The string is capable of pulling only along the length of the string, which in this case is toward the center of the circle. Thus the force exerted by the string is in the direction of the golf ball's acceleration vector. This makes our second example in which the particle's acceleration vector points in the same direction as the force exerted on it.

The example of projectile motion with air resistance, shown in Figure (3), presented a more complex situation. In our study of the motion of a Styrofoam projectile, we had two forces acting on the ball. There was the downward force of gravity, and also the force exerted by the wind we would feel if we were riding along with the ball. We saw that gravity and the wind each produced an acceleration vector, and that the ball's actual accelerations. This is an important clue as to how we should handle situations where more than one force is acting on an object.



#### Figure 3

Gravity and the wind each produce an acceleration,  $\vec{g}$  and  $\vec{a}_{air}$  respectively. The net acceleration of the ball is the vector sum of the two accelerations.

# THE ROLE OF MASS

Our three examples, projectile motion, motion in a circle, and projectile motion with air resistance, all demonstrate that a force produces an acceleration in the direction of the force. The next question is – how much acceleration? Clearly not all forces have the same effect. If I shove a child's toy wagon, the wagon might accelerate rapidly and go flying off. The same shove applied to a Buick automobile will not do very much.

There is clearly a difference between a toy wagon and a Buick. The Buick has much more mass than the wagon, and is much less responsive to my shove.

In our recoil definition of mass discussed in Chapter 6 and illustrated in Figure (4), we defined the ratio of two masses as the inverse ratio of their recoil speeds

$$\frac{\mathbf{m}_1}{\mathbf{m}_2} = \frac{\mathbf{v}_2}{\mathbf{v}_1}$$

The intuitive idea is that the more massive the object, the slower it recoils. The more mass, the less responsive it is to the shove that pushed the carts apart.

Think about the spring that pushes the cart apart in our recoil experiment. Once we burn the thread holding the carts together, the spring pushes out on both carts, causing them to accelerate outward. If the spring is pushing equally hard on both carts (later we will see that it must), then we see that the resulting acceleration and final velocities are inversely proportional to the mass of the cart. If  $m_1$  is twice as massive as  $m_2$ , it gets only half as much acceleration from the same spring force. Our recoil definition and experiments on mass suggests that the effectiveness of a force in producing an acceleration is inversely proportional to the object's mass. For a given force, if you double the mass, you get only half the acceleration. That is the simplest relationship between force and mass that is consistent with our general experience, and it turns out to be the correct one.



**Figure 4** Definition of mass. When two carts recoil from rest, the more massive cart recoils more slowly.

# **NEWTON'S SECOND LAW**

We have seen that a force  $\vec{F}$  acting on a mass m, produces an acceleration  $\vec{a}$  that 1) is in the direction of  $\vec{F}$ , and 2) has a magnitude inversely proportional to m. The simplest equation consistent with these observations is

$$\vec{a} = \frac{\vec{F}}{m}$$
 (1)

Equation (1) turns out to be the correct relationship, and is known as *Newton's Second Law of Mechanics*. (The *First Law* is a statement of the special case that, if there are no forces, there is no acceleration. That was not obvious in the late 1600s, and was therefore stated as a separate law.) A more familiar form of Newton's second law, seen in all introductory physics texts is

$$\vec{F} = m\vec{a}$$
(1a)

If there is any equation that is essentially an icon for the introductory physics course, Equation (1a) is it.

At this point Equation (1) or (1a) serves more as a definition of force than a basic scientific result. We can, for example, see from Equation (1a) that force has the dimensions of mass times acceleration. In the MKS system of units this turns out to be  $kg(m/sec^2)$ , a collection of units called the *newton*. Thus we can say that we push on an object with a force of so many newtons. In the CGS system, the dimensions of force are gm(cm/sec<sup>2</sup>), a set of units called a *dyne*. A dyne turns out to be a very small unit of force, of the order of the force exerted by a fly doing push-ups. The newton is a much more convenient unit. The real confusion is in the English system of units where force is measured in *pounds*, and the unit of mass is a *slug*. We will carefully avoid doing Newton's law calculations in English units so that the student does not have to worry about pounds and slugs.

At a more fundamental level, we can use Equation (1) to *detect the existence of a force by the acceleration it produces*. In projectile motion, how do we know that there is a gravitational force  $\vec{F}_g$  acting on the projectile? Because of the gravitational acceleration. The acceleration  $\vec{a}$  due to gravity is equal to  $\vec{g}$  (9.8 m/sec<sup>2</sup> directed downward), thus we can say that the gravitational force  $\vec{F}_g$  that produces this acceleration is

$$\vec{F}_{g} = m\vec{g}$$
 gravitational force  
on a mass m (2)

where m is the mass of the projectile.



**Figure 5** The gravitational force between small masses is proportional to the product of the masses, and inversely proportional to the square of the separation between them.

# NEWTON'S LAW OF GRAVITY

Newton went beyond using the second law to define force; he also discovered a basic law for the gravitational force between objects. With Newton's law of gravity combined with Newton's second law, we can make detailed predictions about how projectiles, satellites, planets, and solar systems behave. This combination, where one has an explicit formula for gravitational forces, and the second law to predict what accelerations these forces produce, was one of the most revolutionary scientific discoveries ever made.

Newton's so-called *universal law of gravitation* can most simply be stated as follows. If we have two small masses of mass  $m_1$  and  $m_2$ , separated by a distance r as shown in Figure (5), then the force between them is proportional to the product  $m_1m_2$  of their masses, and inversely proportional to the square of the distance r between them. This can be written as an equation of the form

$$\left|\vec{F}_{g}\right| = G \frac{m_{1}m_{2}}{r^{2}}$$
 Newton's law of gravity (3)

where the proportionality constant G is a number that must be determined by experiment.

Equation (3) itself is not the whole story, we must make several more points. First, and very important, is the fact that gravitational forces are always attractive;  $m_1$ is pulled directly toward  $m_2$ , and  $m_2$  directly toward  $m_1$ . Second, the strength of these forces are equal, even if  $m_2$  is much bigger than  $m_1$ , the force of  $m_2$  on  $m_1$  is the same in strength as the force of  $m_1$  on  $m_2$ . That is why we used the same symbol  $\vec{F}_g$  for the two attractive forces in Figure (5).

Newton's law of gravity is called the *universal law of gravitation* because Equation (3) is supposed to apply to all masses anywhere in the universe, with the same numerical constant G everywhere. G is called the *universal gravitational constant*, and has the numerical value, in the MKS system of units

$$G = 6.67 \times 10^{-11} \frac{m^3}{kg \sec^2} \qquad \begin{array}{c} universal \\ gravitational \\ constant \end{array}$$
(4)

We will discuss shortly how this number was first measured.

#### **Exercise 1**

Combine Newton's second law  $\vec{F} = m\vec{a}$  with the law of gravity  $|\vec{F}_g| = Gm_1m_2/r^2$  and show that the dimensions for G in Equation (4) are correct.

### **Big Objects**

In our statement of Newton's law of gravity, we were careful to say that Equation (3) applied to two small objects. To be more explicit, we mean that the two objects  $m_1$  and  $m_2$  should be small in dimensions compared to the separation r between them. We can think of Equation (3) as applying to two *point particles* or *point masses*.

What happens if one or both of the objects are large compared to their separation? Suppose, for example, that you would like to calculate the gravitational force between you and the earth as you stand on the surface of the earth. The correct way to do this is to realize that you are attracted, gravitationally, to every rock, tree, every single piece of matter in the entire earth as indicated in Figure (6). Each of these pieces of matter is pulling on you, and together they produce a net gravitational force  $\vec{F}_g$  which is the force mg that we saw in our discussion of projectile motion.



You are attracted to every piece of matter in the earth.

It appears difficult to add up all the individual forces exerted by every chunk of matter in the entire earth, to get the net force F<sub>g</sub>. Newton also thought that this was difficult, and according to some historical accounts, invented calculus to solve the problem. Even with calculus, it is a fairly complicated problem to add up all of these forces, but the result turns out to be very simple. For any uniformly spherical object, you get the correct answer in Newton's law of gravity if you think of all the mass as being concentrated at a point at the center of the sphere. (This result is an accidental consequence of the fact that gravity is a  $1/r^2$  force, i.e., that it is inversely proportional to the square of the distance. We will have much more to say about this accident in later chapters.)

Since the earth is nearly a uniformly spherical object, you can calculate the gravitational force between you and the earth by treating the earth as a point mass located at its center, 4000 miles below you, as indicated in Figure (7).

#### Galileo's Observation

As we mentioned earlier, Galileo observed that, in the absence of air resistance, all projectiles should have the same acceleration no matter what their mass. This leads to the striking result that, in a vacuum, a steel ball and a feather fall at the same rate. Now we can see that this is a consequence of Newton's second law combined with Newton's law of gravity.



#### Figure 7

The gravitational force of the entire earth acting on you is the same as the force of a point particle with a mass equal to the earth mass, located at the earth's center, one earth radius below you.

Using the results of Figure (7), i.e., calculating  $\vec{F}_{g}$  by replacing the earth by a point mass me located a distance r<sub>e</sub> below us, we get

$$F_{g} = \frac{Gmm_{e}}{r_{e}^{2}}$$
(5)

for the strength of the gravitational force on a particle of mass m at the surface of the earth. Combining this with Newton's second law

$$\vec{F}_g = m\vec{g}$$
 or  $F_g = mg$  (6)

we get

$$mg = \frac{Gm_e}{r^2}$$
(7)

The important result is that the particle's mass m cancels out of Equation (7), and we are left with the formula

$$g = \frac{Gm_e}{r_e^2}$$
(8)

for the acceleration due to gravity. We note that g depends on the earth mass  $m_e$ , the earth radius  $r_e$ , and the universal constant G, but not on the particle's mass *m*. Thus objects of different mass should have the same acceleration.

# Galileo Was Rig

#### By WILLIAM HINES

HOUSTON - Galileo was David R. Scott can prove it.

ematician believed contrary to right. then prevailing scientific opinion that gravity exerts an equal influence on all things, regardless of their size, shape or weight.

He even went so far as to say that if it weren't for the resistance offered by air, a cannonball and a feather would fall at the same speed.

It wasn't possible for Galileo to demonstrate the truth of that assertion, although legend says he performed a compromise experiment by dropping a large iron ball and a small one off the leaning tower of Pisa and that dings."

they hit the ground at the same instant.

Just before getting back into right and Apollo 15 Astronaut the Lunar Landing Craft Falcon fc1 takeoff from the Moon, Scott The 17-th century Italian math- demonstrated that Galileo was

> Holding a metal geological hammer in his right hand and a feather ("A falcon feather," he explained) in his right, Scott faced the Apollo television camera and released the two objects.

> Falling slowly in the eak gravity of the airless Moon, the hammer and the feather reached the surface at precisely the same instant.

> "How about that?" the delighted Scott remarked. "Mr. Galileo was correct in his fin-

# THE CAVENDISH EXPERIMENT

A key feature of Newton's law of gravitation is that all objects attract each other via gravity. Yet in practice, the only gravitational force we ever notice is the force of attraction to the earth. What about the gravitational force between two students sitting beside each other, or between your two fists when you hold them close to each other? The reason that you do not notice these forces is that the gravitational force is incredibly weak, weak compared to other forces that hold you, trees, and rocks together. Gravity is so weak that you would never notice it except for the fact that you are on top of a huge hunk of matter called the earth. The earth mass is so great that, even with the weakness of gravity, the resulting force between you and the earth is big enough to hold you down to the surface.

The gravitational force between two reasonably sized objects is not so small that it cannot be detected, it just requires a very careful experiment that was first performed by Henry Cavendish in 1798. In the Cavendish experiment, two small lead balls are mounted on the end of a light rod. This rod is then suspended on a fine glass fiber as shown in Figure (8a).

As seen in the top view in Figure (8b), two large lead balls are placed near the small ones in such a way that the gravitational force between each pair of large and small balls will cause the rod to rotate in one direction. Once the rod has settled down, the large lead balls are moved to the position shown in Figure (8c). Now the gravitational force causes the rod to rotate the other way. By measuring the angle that the rod rotates, and by measuring what force is required to rotate the rod by this angle, one can experimentally determine the strength of the gravitational force  $\vec{F}_g$  between the balls. Then by using Newton's law of gravity



applied to Figure (9), one can solve for G in terms of the known quantities  $F_g$ ,  $m_1$ ,  $m_2$  and  $r^2$ . This was the way that Newton's universal constant G, given in Equation (4) was first measured.



a) Side view of the small balls.



b) Top view showing two large lead balls.



c) Top view with large balls rotated to new position.

#### Figure 8

The Cavendish experiment. By moving the large lead balls, the small lead balls are first pulled one way, then the other. By measuring the angle the stick holding the small balls is rotated, one can determine the gravitational force  $\vec{F}_{e}$ .

# "Weighing" the Earth

Once you know G, you can go back to the formula (8) for the acceleration g due to gravity, and solve for the earth mass  $m_e$  to get

$$m_{e} = \frac{gr_{e}^{2}}{G} = \frac{9.8 \text{ m/sec}^{2} \times (6.37 \times 10^{6} \text{m})^{2}}{6.67 \times 10^{-11} \text{m}^{3}/\text{kg sec}^{2}}$$
$$= 6.0 \text{ x } 10^{24} \text{kg}$$
(9)

As a result, Cavendish was able to use his value for G to determine the mass of the earth. This was the first determination of the earth's mass, and as a result the Cavendish experiment became known as the experiment that "weighed the earth".

#### **Exercise 2**

The density of water is 1 gram/cm<sup>3</sup>. The average density of the earth's outer crust is about 3 times as great. Use Cavendish's result for the mass of the earth to decide if the entire earth is like the crust. (Hint —the volume of a sphere of radius r is  $4/3\pi r^3$ ). Relate your result to what you have read about the interior of the earth.

# **Inertial and Gravitational Mass**

The fact that, in the absence of air resistance, all projectiles have the same acceleration— the fact that the m's canceled in Equation (7), has a deeper consequence than mere coincidence. In Newton's second law, the m in the formula  $\vec{F} = m\vec{a}$  is the mass defined by the recoil definition of mass discussed in Chapter 6. Called *inertial mass*, it is the concept of mass that we get from the law of conservation of linear momentum.

In Newton's law of gravity, the projectile's mass m in the formula  $F_g = Gmm_e/r_e^2$  is what we should call the *gravitational mass* for it is defined by the gravitational interaction. It is the experimental observation that the m's cancel, the observation that all projectiles have the same acceleration due to gravity, that tells us that the inertial mass is the same as gravitational mass. This equivalence of inertial and gravitational mass has been tested with extreme precision to one part in a billion by Etvös in 1922 and to even greater accuracy by R. H. Dicke in the 1960s.

# SATELLITE MOTION

The key idea that led Newton to his universal law of gravitation was that the moon, while traveling in its orbit about the earth, was subject to the same kind of force as an apple falling from a tree. We have seen that a projectile in flight, such as an apple, accelerates down toward the center of the earth. The moon, in its nearly circular orbit around the earth, also accelerates toward the center of the earth, as illustrated in Figure (10). Newton proposed that the accelerations of the falling apple and of the orbiting moon were both caused by the gravitational pull of the earth.



#### Figure 10

When we swing a golf ball in a circle, the ball accelerates toward the center of the circle, in the direction it is pulled by the string. Similarly, the moon, in its circular orbit about the earth, accelerates toward the center of the earth, in the direction it is pulled by the earth's gravity. The moon, being farther away from the center of the earth should be expected to feel a weaker gravitational force and therefore have a weaker acceleration. From direct calculation Newton could determine how much weaker the moon's acceleration was, and thus determine how the gravitational acceleration and force decreases with distance.

To repeat Newton's calculation, we know that the apple on the surface of the earth has an acceleration  $g_{apple} = 9.8 \text{ m/sec}^2$ . To determine the magnitude of the moon's orbital acceleration toward the earth,  $g_{moon orbit}$ , we can use the formula derived in Chapter 3 for uniform circular motion, namely

$$|\vec{a}| = g_{\text{moon orbit}} = \frac{v^2}{r}$$
 uniform  
circular  
motion (3-12)

To calculate the speed v of the moon, we note that the moon takes 27.32 days or  $2.36 \times 10^6$  seconds for one complete orbit. The radius of the moon orbit is  $3.82 \times 10^8$  meters, so that

$$v_{moon} = \frac{\text{orbital circumference}}{\text{time for one orbit}} = \frac{2\pi r}{t_{\text{orbit}}}$$
$$= \frac{2\pi \times (3.82 \times 10^8 \text{ meters})}{2.36 \times 10^6 \text{ sec}}$$
$$= 1.02 \times 10^3 \frac{\text{m}}{\text{sec}}$$
(10)

or very close to 1 kilometer per second. Substituting this value of v into the formula  $v^2/r$ , gives

$$g_{\text{moon orbit}} = \frac{\left(1.02 \times 10^3 \text{ m/sec}\right)^2}{3.82 \times 10^8 \text{ m}}$$
$$= 2.70 \times 10^{-3} \frac{\text{m}}{\text{sec}^2}$$
(11)

The ratio of the moon's orbital acceleration to the apple's acceleration

$$\frac{g_{\text{moon orbit}}}{g_{\text{apple}}} = \frac{2.70 \times 10^{-3} \text{ m/sec}^2}{9.8 \text{ m/sec}^2}$$
  
= 2.71 × 10<sup>-4</sup> (12)

I.e., the moon's acceleration is 27 thousand times weaker than the apple's.

To understand the meaning of this result, let us look at the square of the ratio of the distances from the apple to the center of the earth, and the moon to the center of the earth. We have

$$\left(\frac{r_{apple to center of earth}}{r_{moon orbit}}\right)^{2} = \left(\frac{6.37 \times 10^{6} \text{ m}}{3.82 \times 10^{8} \text{ m}}\right)^{2}$$
$$= 2.78 \times 10^{-4}$$
(13)

which, to the accuracy of our work, is the same as the ratio of accelerations.

Equating the results in Equations (12) and (13), we get

$$\frac{g_{\text{moon orbit}}}{g_{\text{apple}}} = \frac{r_{e}^{2}}{r_{\text{moon orbit}}^{2}}$$
$$g_{\text{moon orbit}} = \frac{g_{\text{apple}} \times r_{e}^{2}}{r_{\text{moon orbit}}^{2}} \propto \frac{1}{r_{\text{moon orbit}}^{2}}$$
(14)

Where  $g_{apple} \times r_e^2$  can be thought of as a constant.

From such calculations Newton saw that the gravitational acceleration of the moon, and thus the gravitational force, decreased as the square of the distance from the moon to the center of the earth. This was how Newton deduced that gravity was a  $1/r^2$  force law.

#### **Exercise 3**

How far above the surface of the earth do you have to be so that, in free fall, your acceleration is half that of objects near the surface of the earth?

# **Other Satellites**

To explain to the world the similarity of projectile and satellite motion, that both the apple and the moon were simply falling toward the center of the earth, Newton drew the sketch shown in Figure (11). In the sketch, Newton shows a projectile being fired horizontally from the top of a mountain, and shows what would happen if there were no air resistance. If the horizontal velocity were not too great, the projectile would go a short distance along the typical parabolic path we have studied in the strobe labs. As the projectile is fired faster it would travel farther before hitting the ground. Finally we reach a point where the projectile keeps falling toward the earth, but the earth keeps falling away and the projectile goes all the way around the earth without hitting it.

Another perspective of the same idea is illustrated in Figures (12) and (13). Figure (12) is a strobe photograph showing two steel balls launched simultaneously, one being dropped straight down and the other being fired horizontally. The photograph clearly demonstrates that the downward motion of the two projectiles



Newton's sketch showing that the difference between projectile and satellite motion is that satellites travel farther. Both are accelerating toward the center of the earth.

is the same. By using the constant acceleration formulas with g = 32 ft/sec<sup>2</sup>, we can easily calculate that at the end of one second both projectiles will have fallen 16 ft, and at the end of two seconds a distance of 64 ft.

In Figure (13), we have sketched the curved surface of the earth. Due to this curvature, the surface of the earth will be 16 ft below a horizontal line out at a distance of 4.9 miles, and 64 ft below at a distance of 9.8 miles. This effect can be seen from a small boat as you leave shore. When you are 10 miles off shore, you cannot see lighthouses under 64 ft tall, unless you climb your own mast. (For landlubbers sunning on the beach, sailboats with 64 ft high masts disappear from sight at a distance of 10 miles.)

Comparing Figures (12) and (13), we see that in the absence of air resistance, if a projectile were fired horizontally at a speed of 4.9 miles per second, during the first second it would fall 16 ft, but the earth would have also fallen 16 ft, and the projectile would be no closer to the surface. By the end of the 2nd second the projectile would have fallen 64 ft, but still not have come any closer to the surface of the earth. Such a projectile would keep traveling around the earth, never hitting the surface. It would fall all the way around, becoming an earth satellite.



#### Figure 12

Two projectiles, released simultaneously. The horizontal motion has no effect on the vertical motion: they both fall at the same rate.

#### **Exercise 4**

An earth satellite in a low orbit, for instance 100 miles up, is so close to the surface of the earth (100 miles is so small compared to the earth's radius of 4000 miles) that the satellite's acceleration is essentially the same as the acceleration of projectiles here on earth. Use this result to predict the period T of the satellite's orbit. (Hint – the satellite travels one earth circumference  $2\pi r_e$  in one period T. This allows you to calculate the satellite's speed v. You then use the formula v²/r for the magnitude of the satellite's acceleration.)

#### Weight

The popular press often talks about the astronauts in spacecraft orbiting the earth as being *weightless*. This is verified by watching them on television floating around inside the space capsule. You might jump to the conclusion that because the astronauts are floating around in the capsule, they do not feel the effects of gravity. This is true in the same sense that when you jump off a high diving board, you do not feel the effects of gravity—until you hit the water. While you are falling, you are *weightless* just like the astronauts.

The only significant difference between your fall from the high diving board, and the astronaut's weightless experience in the space capsule, is that the astronaut's experience lasts longer. As the space capsule orbits the earth, the capsule and the astronauts inside are in continuous free fall. They have not escaped the earth's gravity, it is gravity that keeps them in orbit, accelerating toward the center of the earth. But because they are in free fall, they do not feel the acceleration, and are considered to be *weightless*.



If the astronaut in an orbiting space capsule is weightless, but still subject to the gravitational force of the earth, we cannot directly associate the word *weight* with the effects of gravity. In order to come up with a definition of the word weight that has some scientific value, and is reasonably consistent with the use of the word in the popular press, we can define the weight of an object as the magnitude of the force the object exerts on the bathroom scales. Here on earth, if you have an object of mass m and you set it on the bathroom scales, it will exert a downward gravitational force of magnitude

$$F_g = mg$$

Thus we say that the object has a weight W given by

$$W = mg \tag{15}$$

For example, a 60 kg boy standing on the scales exerts a gravitational force

$$W(60 \text{ kg boy}) = 60 \text{ kg} \times 9.8 \frac{\text{m}}{\text{sec}^2}$$
$$= 588 \text{ newtons}$$

We see that weight has the dimensions of a force, which in the MKS system is newtons. If the same boy stood on the same scales in an orbiting spacecraft, both the boy and the scales would be in free fall toward the center of the earth, the boy would exert no force on the scales, and he would therefore be weightless.



**Figure 14** We will define the weight of an object as the force it exerts on the bathroom scales.

**Figure 13** *The curvature of the earth causes the horizon to fall away 64 feet at a distance of 9.8 miles.* 

Although we try to make the definition of the word *weight* consistent with the popular use of the word, we do not actually succeed. In almost any country except the United States, when you buy a steak, the butcher will weigh it in grams. The grocer will tell you that a banana weighs 200 grams. You are not likely find a grocer who tells you the weight of an object in newtons. It is a universal convention to tell you the mass in grams or kilograms, and say that that is the weight. About the only place will you will find the word *weight* to mean a force, as measured in newtons, is in a physics course.

(In the English system of units, a pound is a force, so that it is correct to say that our 60 kg mass boy weighs 132 lbs. That, of course, leaves us with the question of what mass is in the English units. From the formula F = mg, we see that m = F/g, or an object that weighs 32 lbs has a mass 32 lbs/32ft/sec<sup>2</sup>=1. As we mentioned earlier, this unit mass in the English units is called a *slug*. This is the last time we will mention slugs in this text.)

### **Earth Tides**

An aspect of Newton's law of gravity that we have not said much about is the fact that gravity is a mutual attraction. As we mentioned, two objects of mass  $m_1$ and  $m_2$  separated by a distance r, attract *each other* with a gravitational force of magnitude  $|\vec{F}_g| = Gm_1m_2/r^2$ . The point we want to emphasize now is that the force on *each* particle has the same strength  $F_g$ .

Let us apply this idea to you, here on the surface of the earth. Explicitly, let us assume that you have just jumped off a high diving board as illustrated in Figure (15), and have not yet hit the water. While you are falling, the earth's gravity exerts a downward force  $\vec{F}_g$  which produces your downward acceleration  $\vec{g}$ .

According to Newton's law of gravity, you are exerting an equal and opposite gravitational force  $\vec{F}_g$  on the earth. Why does nobody talk about this upward force you are exerting on the earth? The answer, shown in the following exercise, is that even though you are pulling up on the earth just as hard as the earth is pulling down on you, the earth is so much more massive that your pull has no detectable effect.

#### **Exercise 5**

Assume that the person in Figure (15) has a mass of 60 kilograms. The gravitational force he exerts on the earth causes an upward acceleration of the earth  $a_{earth}$ . Show that  $a_{earth} = 10^{-22}$  m/sec<sup>2</sup>.



Figure 15 As you fall toward the

As you fall toward the water, the earth is pulling down on you, and you are pulling up on the earth. The two forces are of equal strength. More significant than the force of the diver on the earth is the force of the moon on the earth. It is well known that the ocean tides are caused by the moon's gravity acting on the earth. On the night of a full moon, high tide is around midnight when the moon is directly overhead. The time of high tide changes by about an hour a day in order to stay under the moon.

The high tide under the moon is easily explained by the idea that the moon's gravity sucks the ocean water up into a bulge under the moon. As the earth rotates and we pass under the bulge, we see a high tide. This explains the high tide at midnight on a full moon.

The problem is that there are 2 high tides a day about 12 hours apart. The only way to understand two high tides is to realize that there are two bulges of ocean water, one under the moon and one on the opposite side of the earth, as shown in Figure (16). In one 24 hour period we pass under both bulges.

Why is there a bulge on the backside? Why isn't the water all sucked up into one big bulge underneath the moon?

The answer is that the moon's gravity not only pulls on the earth's water, but on the earth itself. The force of gravity that the moon exerts on the earth is just the same strength as the force the earth exerts on the moon. Since the earth is more massive, the effect on the earth is not as great, but it is noticeable. The reason for the second bulge of water on the far side of the earth is that the center of the earth is closer to the moon than the water on the back side, and therefore accelerates more rapidly toward the moon than the water on the back side. The water on the back side gets left behind to form a bulge.

The result, the fact that there are two high tides a day, the fact that there is a second bulge on the back side, is direct experimental evidence that the earth is accelerating toward the moon. It is direct evidence that the moon's gravity is pulling on the earth, just as the earth's gravity is pulling on the moon.

As a consequence of the earth's acceleration, the moon is not traveling in a circular orbit centered precisely on the center of the earth. Instead both the earth and the moon are traveling in circles about an axis point located on a line joining the earth's and moon's centers. This axis point is located much closer to the center of the earth than that of the moon, in fact it is located inside the earth about 3/4 of the way toward the earth's surface as shown in Figure (17).



**Figure 16** *The two ocean bulges cause two high tides per day.* 





Both the earth and the moon travel in circular orbits about an axis point located about 1/4 of the way down below the earth's surface.

# **Planetary Units**

In introductory physics texts, it has become almost an article of religion that all calculations shall be done using MKS units. This has some advantages - we do not have to talk about pounds and slugs, but practicing physicists seldom follow this rule. Physicists studying the behavior of elementary particles, for example, routinely use a system of units that simplify their calculations, units in which the speed of light and other fundamental constants have the numerical value 1. Using these special units they can quickly solve simple problems and gain an intuitive feeling for which quantities are important and which quantities are not.

In our work with projectiles in the lab the CGS system of units was excellent. The projectiles typically went distances from 10 to 100 cm, in times of the order of 1 second, and had masses of the order of 100 gm. There were no large exponents involved.

Now that we are studying the motion of earth satellites, we are faced with large exponents in quantities like the earth mass and the gravitational constant G which are  $5.98 \times 10^{24}$ kg and  $6.67 \times 10^{-11}$ m<sup>3</sup>/kg sec<sup>2</sup> respectively. The calculations we have done so far using these numbers have required a calculator, and we have had to work hard to gain insight from the results.

Table 1 Planetary Units			
Constant	Symbol	Planetary units	MKS units
Gravitational Constant	G	20	6.67 x $10^{-11}$ m <sup>3</sup> /kg sec <sup>2</sup>
Acceleration due to gravityat the earth's surface	ge	20	9.8 m/sec <sup>2</sup>
Earth mass	m <sub>e</sub>	1	5.98 x 10 <sup>24</sup> kg
Moon mass	m <sub>moon</sub>	.0123	$7.36 \ge 10^{22} \text{ kg}$
Sun mass	m <sub>sun</sub>	3.3 x 10 <sup>5</sup>	1.99 x 10 <sup>30</sup> kg
Metric ton	ton	1.67 x 10 <sup>-22</sup>	1000 kg
Earth radius	r <sub>e</sub>	1	6.37 x 10 <sup>6</sup> m
Moon radius	r <sub>moon</sub>	.2725	1.74 x 10 <sup>6</sup> m
Sun radius	r <sub>sun</sub>	109	6.96 x 10 <sup>8</sup> m
Earth orbit radius	r <sub>earth orbit</sub>	23400	$1.50 \ge 10^{11} \text{ m}$
Moon orbit radius	r <sub>moon orbit</sub>	60	3.82 x 10 <sup>8</sup> m
Hour	hr	1 hr	3600 sec
Moon period	lunar month (siderial)	656 hrs	$2.36 \times 10^{6} \text{ sec}$ (= 27.32 days)
Year	vr	$8.78 \times 10^3$ hrs	$3.16 \times 10^7 \text{ sec}$

We now wish to introduce a new set of units, which we will call *planetary units*, that makes satellite calculations much simpler and more intuitive. One way to design a new set of units is to first decide what will be our unit mass, our unit length, and our unit time, and then work out all the conversion factors so that we can convert a problem into our new units. For working earth satellite problems, we have found that it is convenient to take the earth mass as the unit mass, the earth radius as the unit length, and the hour as the unit time.

$$m_{earth} = 1$$
 earth mass  
 $R_{earth} = 1$  earth radius  
hour = 1

With these choices, speed, for example, is measured in (earth radii)/ hr, etc.

This system of units has a number of advantages. We can set  $m_e$  and  $r_e$  equal to 1 in the gravitational force formulas, greatly simplifying the results. We know immediately that a satellite has crashed if its orbital radius becomes less than 1. Typical satellite periods are a few hours and typical satellite speeds are from 1 to 10 earth radii per hour. What may be a bit surprising is that both the acceleration due to gravity at the surface of the earth, g, and Newton's universal gravitational constant G, have the same numerical value of 20.

Table 1 shows the conversion from MKS to planetary units of common quantities encountered in the study of satellites moving in the vicinity of the earth and the moon.



# A satellite in a low earth orbit.

#### **Exercise 6**

We will have you convert Newton's universal gravitational constant G into planetary units. Start with

$$G = 6.67 \times 10^{-11} \frac{\text{meters}^3}{\text{kg sec}^2}$$

Then multiply or divide by the conversion factors

$$3600 \frac{\text{sec}}{\text{hr}}$$

$$5.98 \times 10^{24} \frac{\text{kg}}{\text{earth mass}}$$

$$6.37 \times 10^{6} \frac{\text{meters}}{\text{earth radii}}$$

until all the dimensions in the formula for G are converted to planetary units. (I.e., convert from seconds to hours, kg to earth mass, and meters to earth radii.) If you do the conversion correctly, you should get the result

$$G = 20 \frac{(\text{earth radii})^3}{(\text{earth mass}) \text{ hr}^2}$$

#### Exercise 7

Explain why g and G have the same numerical value in planetary units.

As an advertisement for how easy it is to use planetary units in satellite calculation, let us repeat Exercise (4) using these units. In that exercise we wished to calculate the period of Sputnik 1, a satellite traveling in a low earth orbit. We were to assume that Sputnik's orbital radius was essentially the earth's radius  $r_e$  as shown in Figure (18), and that Sputnik's acceleration toward the center of the earth was essentially the same as the projectiles we studies in the introductory lab, i.e.,  $g_e = 9.8 \text{ m/sec}^2$ . Using the formula

$$a = \frac{v^2}{r}$$

we get

$$g_e = \frac{v_{Sputnik}^2}{r_e} = \frac{v_{Sputnik}^2}{1}$$

Therefore

$$v_{\text{Sputnik}} = \sqrt{g_e} = \sqrt{20} \frac{\text{earth radii}}{\text{hr}}$$

Now the satellite travels a total distance  $2\pi r_e$  to go one orbit, therefore the time it takes is

Sputnik period = 
$$\frac{2\pi r_e}{v_{\text{Sputnik}}} = \frac{2\pi}{\sqrt{20}} = 1.4 \text{ hrs}$$

Compare the algebra that we just did with what you had to go through to get an answer in Exercise (4). (You should have gotten the same answer, 1.4 hrs, or 84 minutes, or 5,040 seconds. This is in good agreement with the observed time for low orbit satellites.) If you have watched satellite launches on television, you may recall waiting about an hour and a half before the satellite returned.

#### **Exercise 8**

A satellite is placed in a circular orbit whose radius is  $2r_e$  (it is one earth radii above the surface of the earth.)

(a) What is the acceleration due to gravity at this altitude?

(b) What is the period of this satellite's orbit?

(c) What is the shortest possible period any earth satellite can have? Explain your answer.

#### Exercise 9

Communication satellites are usually placed in circular orbits over the equator, at an altitude so that they take precisely 24 hours to orbit the earth. In this way they hover over the same point on the earth and can be in continuous communication with the same transmitters and receivers. This orbit is called the *Clarke orbit*, named after the science fiction writer Arthur Clark who first emphasized the importance of such an orbit. Calculate the radius of the Clark orbit.

# COMPUTER PREDICTION OF SATELLITE ORBITS

In this chapter we have discussed two special kinds of motion that a projectile or satellite can have. One is the parabolic trajectory of a projectile thrown across the room – motion that is easily described by calculus and the constant acceleration formulas. The other is the orbital motion of the moon and man-made satellites that are in circular orbits. These orbits can be analyzed using the fact that their acceleration is directed toward the center of the circle and has a magnitude  $v^2/r$ .

These two examples are deceptively simple. Newton's diagram, Figure (11), shows that there is a continuous range of orbital shapes starting from simple projectile motion out to circular orbital motion and beyond. For all these orbital shapes, we know the projectile's acceleration is the gravitational acceleration toward the center of the earth. But to go from a knowledge of the acceleration to predicting the shape of the orbit is not necessarily an easy task.

There are no simple formulas like the constant acceleration formulas that allow us to predict where the satellite will be at any time in the future. Using advanced calculus techniques one can show that the orbits should have the shape of conic sections, one example being the elliptical orbits discovered by Kepler. But if we go to more complicated problems like trying to predict the motion of the Apollo 8 spacecraft from the earth to the moon and back, then a calculus approach is completely inadequate.

On the other hand these problems are easily handled using the step-by-step method of predicting motion, the method, discussed in Chapter 5, that we implement using the computer. With a slight modification of our old projectile motion program, we can predict what will happen to an earth satellite no matter how it is launched and what orbit it has. Adding a few more lines to the program allows us to send the satellite to the moon and back.

Once we are familiar with a basic satellite motion program, we can easily add new features. We can, for example, change the exponent in the gravitational force law from  $1/r^2$  to  $1/r^{2.1}$  to see what happens if the gravitational force law is modified. Similar modifica-

8-17

tions were in fact predicted by Einstein's general theory of relativity, thus we will be able to observe the kind of effects that were used to verify Einstein's theory.

# **New Calculational Loop**

In Chapter 5, we set up the machinery to do computer calculations. This involved learning the LET statement, constructing loops, plotting crosses, etc. Although this may have been a bit painful (but perhaps not as painful as learning calculus), we do not have to do much of that again. We can use essentially the same machinery to predict satellite orbits. The only significant change is in the calculational loop where we predict the particle's new position and velocity.

In the projectile motion program, the English version of the calculational loop was, from Figure (5-18)

```
! ----- Calculational Loop
DO
```

LET 
$$\vec{R}_{new} = \vec{R}_{old} + \vec{V}_{old} * dt$$
  
LET  $\vec{A} = \vec{g}$   
LET  $\vec{V}_{new} = \vec{V}_{old} + \vec{A} * dt$   
LET  $T_{new} = T_{old} + dt$   
PLOT  $\vec{R}$   
LOOP UNTIL T > 1

#### Figure 19



**Figure 5-15a** *Predicting the next new position.* 

This loop expresses the method of predicting motion that we developed from the analysis of strobe photographs. The idea behind the command

LET 
$$\vec{R}_{new} = \vec{R}_{old} + \vec{V}_{old} * dt$$

is illustrated in Figure (5-15a) reproduced here. The new position of the particle is obtained from the old position by adding the vector  $\vec{V}_{old} * dt$  to the old coordinate vector  $\vec{R}_{old}$ .

Once we get to the new position of the particle, we need the new velocity vector in order to calculate the next new position. The new velocity vector is obtained from the command

LET 
$$\vec{V}_{new} = \vec{V}_{old} + \vec{A} * dt$$

as illustrated in Figure (5-15b). The DO–LOOP part of the program tells us to keep repeating this step-by-step process until we get as much of the trajectory as we want (in this case until one second has elapsed).

The calculational loop of Figure (19) works for projectile motion because we always know the projectile's acceleration  $\vec{A}$  which is given by the line

LET 
$$\vec{A} = \vec{g}$$
 projectile motion (16)

This is the line that characterizes projectile motion, the line that tells the computer that the projectile has a constant acceleration  $\vec{g}$ .



**Figure 5-15b** *Predicting the next new velocity.* 

The only fundamental change we need to make in going from projectile motion to satellite motion is to change our command for the particle's acceleration  $\vec{A}$ . Instead of assuming that the particle's acceleration is constant, we use Newton's law of gravity  $|\vec{F}_g| = Gm_1m_2/r^2$  to calculate the force acting on the satellite, and then Newton's second law  $\vec{A} = \vec{F}_g/m$  to obtain the resulting acceleration.

There are of course some other details. We have to find a way to express the vector nature of the gravitational force -i.e., to tell the computer which way the gravitational force is pointing, and we are going to change our plotting scale since we are no longer working in front of a 100 cm by 100 cm grid. But essentially we are replacing the command

LET 
$$\vec{A} = \vec{g}$$

by the new lines

LET  $\vec{F}_g = GM_e m/R^2$  with instructions for a direction LET  $\vec{A} = \vec{F}_g/m$ 

and then using the same old program.

# **Unit Vectors**

We have no problem describing the direction of the gravitational force on the satellite—the force is directed toward the center of the earth. But how do we tell the computer that? What mathematical technique can we use to express the direction of  $\vec{F}_g$ ?

The technique that we will use throughout the course is the use of the *unit vector*. *A unit vector is a dimensionless vector of length 1 that points in the direction of interest*. If we want a vector of length 5 newtons that points in the same direction, then we multiply our unit vector by the number 5 newtons to get the desired result. (Recall that multiplying a vector by a number, for example n, gives a vector n times as long, pointing in the same direction.)

There is an easy way to construct unit vectors. If we can find some vector that points in the desired direction, we divide that vector by its own length, and we end up with a vector of length 1, the required unit vector. In our satellite motion problem, the gravitational force  $\vec{F}_g$  points toward the center of the earth. Thus to define the direction of  $\vec{F}_g$ , we need a unit vector that points toward the center of the earth. In Figure (20a) we show the coordinate vector  $\vec{R}$  which defines the position of the satellite in a coordinate system whose origin is at the center of the earth. In Figure (20b) we see the vector  $-\vec{R}$ , which points from the satellite to the center of the earth, the same direction as the gravitational force. Therefore we would like to turn  $-\vec{R}$  into a unit vector, which we do by dividing by the length of R, namely the distance from the center of the earth to the satellite.

Since we will often use unit vectors in this text, we will designate them by a special symbol. Instead of an arrow over the letter, we will use what is called a *caret* by typographers, or more familiarly a *hat* by physicists. Thus our unit vector in the  $-\vec{R}$  direction will be denoted by  $-\hat{R}$  and is given by the formula



In Equation (16), the length R is given by the Pythagorean theorem

$$\mathbf{R} = \sqrt{\mathbf{R}_{\mathrm{x}}^2 + \mathbf{R}_{\mathrm{y}}^2} \tag{16a}$$

 $R_x$  and  $R_y$  being the x and y coordinates of the satellite.

With the unit vector  $-\hat{R}$ , we can now write an explicit formula for the gravitational force vector  $\vec{F}_g$ . We multiply the unit vector  $-\hat{R}$  by the magnitude GMm/R<sup>2</sup> of the gravitational force to get

$$\vec{F}_{g} = \frac{GM_{e}m}{R^{2}} (-\hat{R})$$
(17)

# Calculational Loop for Satellite Motion

We are now ready to go in an orderly way from the calculational loop for projectile motion to a calculational loop for satellite motion. We can focus our attention on the following three lines of the projectile motion calculation loop (Figure 21) because the other lines remain unchanged.

LET 
$$\vec{R}_{new} = \vec{R}_{old} + \vec{V}_{old} * dt$$
  
LET  $\vec{A} = \vec{g}$   
LET  $\vec{V}_{new} = \vec{V}_{old} + \vec{A} * dt$ 

#### Figure 21

The first step is to replace LET  $\vec{A} = \vec{g}$  by Newton's law of gravity and Newton's second law as shown in Figure (22).

LET 
$$\vec{R}_{new} = \vec{R}_{old} + \vec{V}_{old} * dt$$
  
LET  $\vec{F}_g = (-\hat{R})GM_em/R^2$   
LET  $\vec{A} = \vec{F}_g/m$   
LET  $\vec{V}_{new} = \vec{V}_{old} + \vec{A} * dt$ 

#### Figure 22

Because BASIC is limited to working with numerical commands rather than vectors (an unfortunate limitation), the next step is to make sure that we can translate each of these vector commands into the separate x and y components. We will do this separately for each of the 4 lines.

The command

LET 
$$\vec{R}_{new} = \vec{R}_{old} + \vec{V}_{old} * dt$$

for  $\vec{R}_{new}$  becomes

$$LET Rx = Rx + Vx * dt$$
(18a)

$$LET Ry = Ry + Vy * dt$$
(18b)

where we drop the subscripts "new" and "old" because the computer automatically takes the old values on the right side of the LET statement, calculates a new value, and stores the new value in the memory cell named on the left side of the LET statement. (See our discussion of the LET statement on page 5-5).

In Equations (18a) and (18b) we obtain numerical values for the new coordinates Rx and Ry of the satellite. However, we will also need to know the distance R from the satellite to the center of the earth (in order to construct the unit vector  $-\hat{R}$ ). The value of R is easily determined by adding the command

$$LET R = SQR (Rx*Rx + Ry*Ry)$$
(18c)

where SQR is BASIC's way of saying square root.

The translation of the command for  $\vec{F}_g$  only requires the translation of the unit vector  $\hat{R}$  into x and y coordinates. Remembering that  $\hat{R} = \vec{R}/R$ , we get

$$\widehat{R}_{x} = Rx/R; \qquad \widehat{R}_{y} = Ry/R$$
 (19)

thus the translation of the LET statement for  $\vec{F}_g\,$  can be written as

LET Fg = 
$$G * Me * M / (R*R)$$
  
LET Fgx =  $(-Rx / R) * Fg$   
LET Fgy =  $(-Ry / R) * Fg$ 

The computer can handle these lines because it already knows the new values of Rx, Ry and R from Equations (18a, b, and c).

The translation of LET statements for  $\vec{A}$  and  $\vec{V}_{new}$  are straightforward. We get

We included a calculation of the magnitude V of the satellite's speed for future use. We may, for example, want to construct a unit vector in the  $-\vec{V}$  direction to represent the direction of air resistance on a reentering satellite. We have found it convenient to routinely calculate the magnitude of any vector whose x and y coordinates we have just calculated.

# Summary

To summarize our translation, we started with the vector commands

LET 
$$\vec{R}_{new} = \vec{R}_{old} + \vec{V}_{old} * dt$$
  
LET  $\vec{F}_g = (-\hat{R})GM_em/R^2$   
LET  $\vec{A} = \vec{F}_g/m$   
LET  $\vec{V}_{new} = \vec{V}_{old} + \vec{A} * dt$ 

and ended up with the BASIC commands

LET 
$$Rx = Rx + Vx * dt$$
  
LET  $Ry = Ry + Vy * dt$   
LET  $R = SQR (Rx*Rx + Ry*Ry)$   
LET  $Fg = G * Me * M / (R*R)$   
LET  $Fgx = (-Rx / R) * Fg$   
LET  $Fgy = (-Ry / R) * Fg$   
LET  $Ax = Fgx / M$   
LET  $Ay = Fgy / M$   
LET  $Vx = Vx + Ax * dt$   
LET  $Vy = Vy + Ay * dt$   
LET  $V = SQR(Vx*Vx + Vy*Vy)$ 

# **Working Orbit Program**

We are now ready to convert a working projectile motion program, Figure (5-23) reproduced here, into a working orbital motion program. In addition to converting the calculational loop as we have just discussed, we need to change some constants and plotting ranges, but the general structure of the program will be unchanged.

#### **Plotting Window**

We will initially consider satellite motion that stays reasonably close to the earth, within several earth radii. Using planetary units, and placing the earth at the center of the plot, we can get a reasonable range of orbits if we let Rx vary for example from - 9 to +9 earth radii. If we have a standard 9" Macintosh screen, the x dimension should be 1.5 times the y dimension, thus Ry should go only from -6 to +6. The following command sets up this plotting window

# SET WINDOW -9, 9, -6, 6

To show where the earth is located, we can use the following lines to plot a cross at the center of the earth

# **Constants and Initial Conditions**

In going from the projectile motion to the satellite motion program, we have to change the constants and initial conditions. Using planetary units, our constants G, Me, and m are

LET G = 
$$20$$
  
LET Me =  $1$   
LET m = .001

(Our choice of the satellite mass m does not matter because it cancels out of the calculation.)

For initial conditions, we will start the satellite .1 earth radii above the surface of the earth on the + x axis;

LET Rx = 1.1LET Ry = 0LET R = SQR(Rx\*Rx + Ry\*Ry)CALL CROSS

#### **Projectile Motion Program**

```
! ----- Plotting window
!
           (x axis = 1.5 times y axis)
   SET WINDOW -40,140,-10,110
! ----- Draw & label axes
   BOX LINES 0,100,0,100
   PLOT TEXT, AT -3,0 : "O"
   PLOT TEXT, AT -13,96: "y=100"
   PLOT TEXT, AT 101,0 : "x=100"
! ----- Initial conditions
  LET DeltaT = .1
  LET R_{X} = 25.9
  LET Ry = 89.9
  LET U_X = (43.2 - 8.3)/(2*.1)
  LET Vy = (90.2 - 79.3)/(2*.1)
  LET T = 0
   CALL CROSS
! ----- Computer Time Step
  LET dt = .001
  LET i = 0
! ----- Calculational loop
   DO
     LET Rx = Rx + Vx*dt
     LET Ry = Ry + Vy*dt
     LET A \times = 0
     LET Ay = -980
     LET Vx = Vx + Ax*dt
     LET Vy = Vy + Ay*dt
     LET T = T + dt
     LET i = i+1
     IF MOD(i,100) = 0 THEN CALL CROSS
      PLOT Rx, Ry
   LOOP UNTIL RX > 100
! ----- Subroutine "CROSS" draws
     ! a cross at Rx,Ry.
   SUB CROSS
      PLOT LINES: Rx-2, Ry; Rx+2, Ry
       PLOT LINES: Rx, Ry-2; Rx, Ry+2
   END SUB
END
```

#### Figure 23

Projectile motion program that plots crosses every tenth of a second.

#### **Orbit-1 Program**

```
!----- Plotting window
!
        (x axis = 1.5 times y axis))
  SET WINDOW -9,9,-6,6
!----- Plot cross at center of the Earth
   LET R_X = 0
    LET Ry = 0
    CALL CROSS
!----- Constants and Initial Conditions
  LET G = 19.91
  LET Me = 1
  LET M = .001
  LET R_X = 1.1
  LET Ry = O
  LET R = SQR(R_X * R_X + R_q * R_q)
  CALL CROSS
  LET V_X = 0
  LET Vy = 5.5
  LET T = 0
!----- Time step
  LET dt = .01
  LET i = 0
!----- Caculational loop
   DO
     LET Rx = Rx + Vx*dt
     LET Ry = Ry + Vy*dt
     LET R = SQR(R_{x}*R_{x} + R_{y}*R_{y})
     LET F = G*M*Me/R^2
     LET F \times = F * (-R \times / R)
     LET Fy = F*(-Ry/R)
     LET Ax = Fx/M
     LET Ay = Fy/M
     LET Ux = Ux + Ax*dt
      LET Vy = Vy + Ay*dt
     LET U = SQR(Ux*Ux + Uy*Uy)
     LET T = T + dt
      LET i = i + 1
      PLOT Rx, Ry
      IF MOD(i,40) = 0 THEN CALL CROSS
   LOOP UNTIL T > 8.45
   !----- Subroutine "CROSS" draws
           ! a cross at Rx,Ry.
   SUB CROSS
      PLOT LINES: Rx-.2, Ry; Rx+.2, Ry
      PLOT LINES: Rx, Ry-.2; Rx, Ry+.2
   END SUB
END
```

#### **Figure 24** *Our new orbital motion pr*

Our new orbital motion program.

We also calculated an initial value of R for use in the gravitational force formula, and plotted a cross at this initial point.

We are going to fire the satellite in the +y direction, parallel to the surface of the earth. Trial and error shows us that a reasonable value for the speed of the satellite is 5.5 earth radii per hour, thus we write for our initial velocity commands

LET 
$$Vx = 0$$
  
LET  $Vy = 5.5$   
LET  $V = SQR(Vx*Vx + Vy*Vy)$ 

In our projectile motion program of Figure (5-23) we wanted a cross plotted every 100 time steps dt. This was done with the command

IF MOD(i,100) = 0 THEN CALL CROSS

For our orbit program, trial and error shows that we get a good looking plot if we draw a cross every 40 time steps, each time step dt being .01 hours. Thus our new MOD line will be

IF MOD(i,40) = 0 THEN CALL CROSS

and we will get a cross every .01 \* 40 = .4 hours.

The final change is to stop plotting after one orbit. From running the program we find that one orbit takes about 9 hours, thus we can stop plotting just before one orbit with the LOOP instruction

LOOP UNTIL T > 9

Putting all these steps together gives us the complete BASIC program shown in Figure (24).

When we run the Orbit 1 program, we get the elliptical orbit shown in Figure (25).

# Exercise 10

Convert your projectile motion program to the Orbit 1 program. Use the same initial conditions so that you get the same orbit as that shown in Figure (25). (It is important to get your Orbit 1 program running correctly now, for it will be used as the basis for studying several phenomena during the rest of this chapter. If you are having problems, simply type the program in precisely as shown in Figure (24).





Output of the Orbit 1 program. The satellite is initially out at a distance x = 1.1 earth radii, and is fired in the +y direction at a speed of 5.5 earth radii per hour.

Once your program is working, it is easy to make small modifications to improve the results. To create Figure (25a) we added the command

BOX CIRCLE -1,1,-1,1

to draw a circle to represent the earth. We also changed dt to .001 and changed the MOD command to MOD(i,539) to get an even number of crosses around the orbit. We then plotted until T = 9 hours. (With dt ten times smaller, our i counter has to be ten times bigger to get the old crosses.)



Figure 25a

#### **Satellite Motion Laboratory**

In our study of projectile motion, we could go to the laboratory and take strobe photographs in order to see how projectiles behaved. Obtaining experimental data for the study of satellite motion is somewhat more difficult. What we will do is to use the Orbit 1 program or slight modification of it to stimulate satellite motion, using it as our laboratory for the study of the behavior of satellites.

But first we wish to check that the Orbit 1 program makes predictions that are in agreement with experiment. The program is based on Newton's laws of gravity,  $F_g = GMm/r^2$ , Newton's law of motion  $\vec{a} = \vec{F}/m$ , and the procedures we developed earlier for predicting the motion of an object whose acceleration is known. Thus a verification of the results of the Orbit 1 program can be considered a verification of these laws and procedures.

Some tests of the Orbit 1 program can be made using the results of your own experience. Anyone who has listened to the launch of a low orbit satellite should be aware that the satellite takes about 90 minutes to go around the earth once. The Orbit 1 program should give the same result, which you can check in Exercise 11. Another obvious test is the prediction of the period of the moon in its orbit around the earth. It is about 4 weeks from full moon to full moon, thus the period should be approximately 4 weeks or 28 days. The fact that the apparent diameter of the moon does not change much during this time indicates that the moon is traveling in a nearly circular orbit about the earth. If you accept the astronomer's measurements that the moon orbit radius is about 60 earth radii away, then you can check the Orbit 1 program to see if it predicts a 4 week period for an earth satellite in a circular orbit of that radius (Exercise 12).

(An easy way to measure the distance to the moon was provided by the first moon landing. Because of a problem with Neil Armstrong's helmet, radio signals sent to Neil from Houston were retransmitted by Neil's microphone, giving an apparent echo. The echo was particularly noticeable while Neil was setting up a TV camera. On a tape of the mission supplied by NASA, you can hear the statement "That's good there, Neil". A short while later you hear the clear echo "That's good there, Neil". The time delay from the original statement and the echo is the time it takes a radio wave, traveling at the speed of light, to go to the moon and back. Using an inexpensive stop watch, one can easily measure the time delay as being about 2 2/5 seconds. Thus the one-way trip to the moon is 1 1/5 seconds. Since light travels 1 ft/nanosecond, or 1 billion feet per second, from this one determines that the moon is about 1.2 billion feet away. You can convert this distance to earth radii to check the astronomer's value of 60 earth radii as the average distance to the moon.)

#### Exercise 11

Adjust the initial conditions in your Orbit 1 program so that the satellite is in a low earth orbit, and see what the period of the orbit is.

(To adjust the initial conditions, start, for example, with  $R_x = 1.01$ ,  $R_y = 0$ ,  $v_x = 0$  and adjust  $v_y$  until you get a circular orbit centered on the earth. As a check that the satellite did not go below the surface of the earth, you could add the line

#### IF R < 1 THEN PRINT "CRASHED"

Adding this line just after you have calculated R in the DO LOOP will immediately warn you if the satellite has crashed. You can then adjust the initial  $v_y$  so that you just avoid a crash. Once you have a circular orbit, you can adjust the time in the "LOOP UNTIL T > ..." command so that just one orbit is printed. This tells you how long the orbit took. You can also see how long the orbit took by adding the line in the DO LOOP

IF MOD(I, 40) = 0 THEN PRINT T, RX, RY

Looking at the values of  $R_x$  and  $R_y$  you can tell when one orbit is completed, and the value of T tells you how long it took.

#### Exercise 12

Put the satellite in a circular orbit whose radius is equal to the radius of the moon's orbit. (See Table 1, Planetary Units, for the value of the moon orbit radius.) See if you predict that the moon will take about 4 weeks to go around this orbit.)

# **KEPLER'S LAWS**

A more detailed test of Newton's laws and the Orbit 1 program is provided by Kepler's laws of planetary motion.

To get a feeling for the problems involved in studying planetary motion, imagine that you were given the job of going outside, looking at the sky, and figuring out how celestial objects moved. The easiest to start with is the moon, which becomes full again every four weeks. On closer observation you would notice that the moon moved past the background of the apparently fixed stars, returning to its original position in the sky every 27.3 days. Since, as we mentioned, the diameter of the moon does not change much, you might then conclude that the moon is in a circular orbit about the earth, with a period of 27.3 days.

The time it takes the moon to return to the same point in the sky is not precisely equal to the time between full moons. A full moon occurs when the sun, earth, and moon are in alignment. If the sun itself appears to move relative to the fixed stars, the full moons will not occur at precisely the same point, and the time between full moons will not be exactly the time it takes the moon to go around once.

To study the motion of the sun past the background of the fixed stars is more difficult because the stars are not visible when the sun is up. One way to locate the position of the sun is to observe what stars are overhead at "true" midnight, half way between dusk and dawn. The sun should then be located on the opposite side of the sky. (You also have to correct for the north/south position of the sun.) After a fair amount of observation and calculations, you would find that the sun itself moves past the background of the fixed stars, returning to its starting point once a year.

From the fact that the sun takes one year to go around the sky, and the fact that its apparent diameter remains essentially constant, you might well conclude that the sun, like the moon, is traveling in a circular orbit about the earth. This was the accepted conclusion by most astronomers up to the time of Nicolaus Copernicus in the early 1500s AD.

If you start looking at the motion of the planets like Mercury, Venus, Mars, Jupiter, and Saturn, all easily visible without a telescope, the situation is more complicated. Mars, for example, moves in one direction against the background of the fixed stars, then reverses and goes backward for a while, then forward again as shown in Figure (26). None of the planets has the simple uniform motion seen in the case of the moon and the sun.

After a lot of observation and the construction of many plots, you might make a rather significant discovery. You might find what the early Greek astronomers learned, namely that if you assume that the planets Mercury, Venus, Mars, Jupiter, and Saturn travel in circular orbits about the sun, while the sun is traveling in a circular orbit about the earth, then you can explain all the peculiar motion of the planets. This is a remarkable simplification and compelling evidence that there is a simple order underlying the motion of celestial objects.

One of the features of astronomical observations is that they become more accurate as time passes. If you observe the moon for 100 orbits, you can determine the average period of the moon nearly 100 times more accurately than from the observation of a single period. You can also detect any gradual shift of the orbit 100 times more accurately.



Retrograde motion of the planet Mars. Modern view of why Mars appears to reverse its direction of motion for a while.

Even by the time of the famous Greek astronomer Ptolemy in the second century AD, observations of the positions of the planets had been made for a sufficiently long time that it had become clear that the planets did not travel in precisely circular orbits about the sun. Some way was needed to explain the non circularity of the orbits.

The simplicity of a circular orbit was such a compelling idea that it was not abandoned. Recall that the apparently peculiar motion of Mars could be explained by assuming that Mars traveled in a circular orbit about the sun which in turn traveled in a circular orbit about the earth. By having circular orbits centered on points that are themselves in circular orbits, you can construct complex orbits. By choosing enough circles with the correct radii and periods, you can construct any kind of orbit you wish.

Ptolemy explained the slight variations in the planetary orbits by assuming that the planets traveled in circles around points which traveled in circles about the sun, which in turn traveled in a circle about the earth. The extra cycle in this scheme was called an *epicycle*. With just a few epicycles, Ptolemy was able to accurately explain all observations of planetary motion made by the second century AD.

With 1500 more years of planetary observations, Ptolemy's scheme was no longer working well. With far more accurate observations over this long span of time, it was necessary to introduce many more epicycles into Ptolemy's scheme in order to explain the positions of the planets.

Even before problems with Ptolemy's scheme became apparent, there were those who argued that the scheme would be simpler if the sun were at the center of the solar system and all the planets, including the earth, moved in circles about the sun. This view was not taken seriously in ancient times, because such a scheme would predict that the earth was moving at a tremendous speed, a motion that surely would be felt. (The principle of relativity was not understood at that time.)

For similar reasons, one did not use the rotation of the earth to explain the daily motion of sun, moon, and stars. That would imply that the surface of the earth at the equator would be moving at a speed of around a thousand miles per hour, an unimaginable speed!

In 1543, Nicolaus Copernicus put forth a detailed plan for the motion of the planets from the point of view that the sun was the center of the solar system and that all the planets moved in circular orbits about the sun. Such a theory not only conflicted with common sense about feeling the motion of the earth, but also displaced the earth and mankind from the center of the universe, two results quite unacceptable to many scholars and theologians.

Copernicus' theory was not quite as simple as it first sounds. Because of the accuracy with which planetary motion was know by 1543, it was necessary to include epicycles in the planetary orbits in Copernicus' model.

Starting around 1576, the Dutch astronomer Tycho Brahe made a series of observations of the planetary positions that were a significant improvement over previous measurements. This work was done before the invention of the telescope, using apparatus like that shown in Figure (27). Tycho Brahe did not happen to believe in the Copernican sun-centered theory, but that had little



**Figure 27** *Tycho Brahe's apparatus.* 

effect on the reason for making the more accurate observations. Both the Ptolemaic and Copernican systems relied on epicycles, and more accurate data was needed to improve the predictive power of these theories.

Johannes Kepler, a student of Tycho Brahe, started from the simplicity inherent in the Copernican system, but went one step farther than Copernicus. Abandoning the idea that planetary motion had to be described in terms of circular orbits and epicycles, Kepler used Tycho Brahe's accurate data to look for a better way to describe the planet's motion. Kepler found that the planetary orbits were accurately and simply described by ellipses, where the sun was at one of the focuses of the ellipse. (We will soon discuss the properties of ellipses.) Kepler also found a simple rule relating the speed of the planet to the area swept out by a line drawn from the planet to the sun. And thirdly, he discovered that the ratio of the cube of the orbital radius to the square of the period was the same for all planets. These three results are known as Kepler's three laws of planetary motion.

Kepler's three simple rules for planetary motion, which we will discuss in more detail shortly, replaced and improved upon the complex system of epicycles needed by all previous theories. After Kepler's discovery, it was obvious that the sun-centered system and elliptical orbits provided by far the simplest description of the motion of the heavenly objects. For Isaac Newton, half a century later, Kepler's laws served as a fundamental test of his theories of motion and gravitation. We will now use Kepler's laws in a similar way, as a test of the validity of the Orbit 1 program and our techniques for predicting motion.



**Figure 28** *Ellipse constructed with two nails and a string.* 

# **Kepler's First Law**

Kepler's first law states that the planets move in elliptical orbits with the sun at one focus. By analogy we should find from our Orbit 1 program that earth satellites move in elliptical orbits with the center of the earth at one focus. To check this prediction, we need to know how to construct an ellipse and determine where the focus is located.

The arch above the entrance to many of the old New England horse sheds was a section of an ellipse. The carpenters drew the curve by placing two nails on a wide board, attaching the ends of a string to each nail, and moving a pencil around while keeping the string taut as shown in Figure (28). The result is half an ellipse with a nail at each one of the focuses. (If you are in the Mormon Tabernacle's elliptical auditorium and drop a pin at one focus, the pin drop can be heard at the other focus because the sound waves bouncing off the walls all travel the same distance and add up constructively at the second focus point.)

To see if the satellite orbit from the Orbit 1 program is an ellipse, we first locate the second focus using the output shown in Figure (25a) by locating the point symmetrically across from the center of the earth as shown in Figure (29). Then at several points along the orbit we draw lines from that point to each focus as shown, and see if the total length of the lines (what would be the length of the stretched string) remains constant as we go around the orbit).





Checking that our satellite orbits are an ellipse. We construct a second focus, and then see if the sum of the distances from each focus to a point on the ellipse in the same for any point around the ellipse. For this diagram, we should show that a+b = c+d.

#### Exercise 13

Using the output from your Orbit 1 program, check that the orbit is an ellipse.

#### Exercise 14

Slightly alter the initial conditions of your Orbit 1 program to get a different shaped orbit. (Preferably, make the orbit more stretched out.) Check that the resulting orbit is still an ellipse.

# **Kepler's Second Law**

Kepler's second law relates the speed of the planet to the area swept out by a line connecting the sun to the planet. If we think of the sun as being at the origin of the coordinate system, then the line from the sun to the planet is what we have been calling the *coordinate vector*  $\vec{R}$ . It is also called the *radius vector*  $\vec{R}$ . Kepler's second law explicitly states *that the radius vector*  $\vec{R}$  sweeps out equal areas in equal times.

To apply Kepler's second law to the output of our Orbit 1 program, we note that we had the computer plot a cross at equal times along the orbit. Thus the area swept out by the radius vector should be the same as  $\vec{R}$  moves from one cross to the next. To check this prediction, we

have in Figure (30) reproduced the output of Figure (25a), shaded the areas swept out as  $\vec{R}$  moves from positions A to B, from C to D, and from E to F. These areas should look approximately equal; you will check that they are in fact equal in Exercise 15.

The most significant consequence of Kepler's second law is that in order to sweep out equal areas while the radius vector is changing length, the planet or satellite must move more rapidly when the radius vector is short, and more slowly when the radius vector is long. The planet moves more rapidly when in close to the sun, and more slowly when far away.

An extreme example of elliptical satellite orbits are the orbits of some of the comets that periodically visit the sun. Halley's comet, for example, visits the sun once every 76 years. The comet spends about 1 year in the close vicinity of the sun, where it is visible from the earth, and the other 75 years on the rest of its orbit which goes out beyond the edge of the planetary system. The comet moves rapidly past the sun, and spends the majority of the 76 year orbital period creeping around the back side of its orbit where its radius vector is very long.



Kepler's Second Law. The radius vector  $\vec{R}$  should sweep out equal areas in equal time.

#### **Exercise 15**

For both of your plots from Exercises 13 and 14, check that the satellite's radius vector sweeps out equal areas in equal times. Explicitly compare the area swept out during a time interval where the satellite is in close to the earth to an equal time interval where the satellite is far from the earth.

This exercise requires that you measure the areas of lopsided pie-shaped sections. There are a number of ways of doing this. You can, for example, draw the sections out on graph paper and count the squares, you can break the areas up into triangles and calculate the areas of the triangles, or you can cut the areas out of cardboard and weigh them.

# **Kepler's Third Law**

Kepler's third law states that the ratio of the cube of the orbital radius R to the square of the period T is the same ratio for all the planets. We can easily use Newton's laws of gravity and motion to check this result for the case of circular orbits. The result, which you are to calculate in Exercise 16, is

$$\frac{\mathrm{R}^3}{\mathrm{T}^2} = \frac{\mathrm{GM}_{\mathrm{s}}}{4\pi^2} \tag{20}$$

where  $M_s$  is the mass of the sun. In this calculation, the mass  $m_p$  of the planet, the orbital radius R, the speed v all cancelled, leaving only the sun mass  $M_s$  as a variable. Since all the planets orbit the same sun, this ratio should be the same for all the planets.

When the planet is in an elliptical orbit, the length of the radius vector  $\hat{R}$  changes as the planet goes around the sun. What Kepler found was that the ratio of  $R^3/T^2$  was constant if you used the "semi major axis" for R. The semi major axis is the half the maximum diameter of the ellipse, shown in Figure (31). As an optional Exercise (17), you can compare the ratio of  $R^3/T^2$  for the two elliptical orbits of Exercises (13) and (14), using the semi major axis for R.

#### **Exercise 16**

Consider the example of a planet of mass  $m_p$  in a circular orbit about the sun whose mass is  $M_g$ . Using Newton's second law and Newton's law of gravity, and the fact that for circular motion the magnitude of the acceleration is  $v^2/R$ , solve for the radius R of the orbit. Then use the fact that the period T is the distance  $2\pi R$  divided by the speed v, and construct the ratio  $R^3/T^2$ . All the variables except  $M_g$  should cancel and you should get the result shown in Equation 20.

#### **Exercise 17 (optional)**

A more general statement of Kepler's third law, that applies to elliptical orbits, is that  $R^3/T^2$  is the same for all the planets, where R is the semi major axis of the ellipse (as shown in Figure (31)). Check this prediction for the two elliptical orbits used in Exercises (13) and (14). In both of those examples the satellite was orbiting the same earth, thus the ratios should be the same.



**Figure 31** *The semi major axis of an ellipse.* 

# MODIFIED GRAVITY AND GENERAL RELATIVITY

After we have verified that the Orbit 1 program calculates orbits that are in agreement with Kepler's laws of motion, we should be reasonably confident that the program is ready to serve as a laboratory for the study of new phenomena we have not necessarily encountered before. To illustrate what we can do, we will begin with a question that cannot be answered in the lab. What would happen if we modified the law of gravity? What, for example, would happen if we changed the universal constant G, or altered the exponent on the r dependence of the force? With the computer program, these questions are easily answered. We simply make the change and see what happens.

These changes should not be made completely without thought. I have seen a project where a student tried to observe the effect of changing the mass of the satellite. After many plots, he concluded that the effect was not great. That is not a surprising result considering the fact that the mass  $m_s$  of the satellite cancels out when you equate the gravitational force to  $m_s \vec{a}$ .

One can also see that, as far as its effect on a satellite's orbit, changing the universal constant G will have an effect equivalent to changing the earth mass  $M_e$ . Since Kepler's laws did not depend particularly on what mass our sun had, one suspects that Kepler's laws should also hold when G or  $M_e$  are modified. This guess can easily be checked using the Orbit 1 program.

Changing the r dependence of the gravitational force is another matter. After developing the special theory of relativity, Einstein took a look at Newton's theory of gravity and saw that it was not consistent with the principle of relativity. For one thing, because the Newtonian gravitational force is supposed to point to the current instantaneous position of a mass, it should be possible using Newtonian gravity to send signals faster than the speed of light. (Think about how you might do that.)

From the period of time between 1905 and 1915 Einstein worked out a new theory of gravity that was consistent with special relativity and, in the limit of slowly moving, not too massive objects, gave the same results as Newtonian gravity. We will get to see how this process works when, in the latter half of this text we start with Coulomb's electric force law, include the effects of special relativity, and find that magnetism is one of the essential consequences of this combination.

Einstein's relativistic theory of gravity is more complex than the theory of electricity and magnetism, and the new predictions of the theory are much harder to test. It turns out that Newtonian gravity accurately describes almost all planetary motion we can observe in our solar system. Einstein calculated that his new theory of gravity should predict new observable effects only in the case of the orbit of Mercury and in the deflection of starlight as it passed the rim of the sun. In 1917 Sir Arthur Eddington led a famous eclipse expedition in which the deflection of starlight past the rim of the eclipsed sun could be observed. The deflection predicted by Einstein was observed, making this the first clear correction to Newtonian gravity detected in 250 years. Einstein's real fame began with the success of the Eddington expedition.

While Einstein set out to construct a theory of gravity consistent with special relativity, he was also impressed by the connection between gravity and space. Because all projectiles here on the surface of the earth have the same downward acceleration, if you were in a sealed room you could not be completely sure whether your room was on the surface of the earth, and the downward accelerations were caused by gravity, or whether you were out in space, and your room was accelerating upward with an acceleration g. These equivalent situations are shown in Figure (32).



#### Figure 32

Equivalent situations. Explain why you would feel the same forces if you were sitting on the floor of each of the two rooms.
The equivalence between a gravitational force and an acceleration turned out to be the cornerstone of Einstein's relativistic theory of gravity. It turned out that Einstein's new theory of gravity could be interpreted as a theory of space and time, where mass caused a curvature of space, and what we call gravitational forces were a consequence of this curvature of space. This geometrical theory of gravity, Einstein's relativistic theory, is commonly called the *General Theory of Relativity*.

As they often say in textbooks, a full discussion of Einstein's relativistic theory of gravity is "beyond the scope of this text". However we can look at at least one of the predictions. As far as satellite orbit calculations are concerned, we can think of Einstein's theory as a slight modification of the Newtonian theory. We have seen that any modification of the factors G,  $m_s$  or  $m_e$  in the Newtonian gravitational force law would not have a detectable effect. The only thing we could notice is some change in the exponent of r.

With a few of quick runs of the Orbit 1 program, you will discover that the satellite orbit is very sensitive to the exponent of r. In Figure (33) we have changed the exponent from -2 to -1.9. This simply requires changing

$$G*m_s*m_e/(R_A2)$$
  
to  
 $G*m_s*m_e/(R_A1.9)$ 

in the formula for  $F_g$ . The result is a striking change in the orbit. When the exponent is – 2, the elliptical orbit is rock steady. When we change the exponent to – 1.9, the ellipse starts rotating around the earth. This rotation of the ellipse is called the *precession of the perihelion*, where the word "perihelion" describes the line connecting the two focuses of the ellipse. A  $1/r^2$  force law is unique in that only for this exponent, -2, does the perihelion, the axis of the elliptical orbit, remain steady. For any other value of the exponent, the perihelion rotates or precesses one way or another.

It turns out that a number of effects can cause the perihelion of a planet's orbit to precess. The biggest effect we have not yet discussed is the fact that there are a number of planets all orbiting the sun at the same time, and these planets all exert slight forces on each other. These slight forces cause slight perihelion precessions.

In the 250 years from the time of Newton's discovery of the law of gravity, to the early 1900s, astronomers carefully worked out the predicted orbits of the planets, including the effects of the forces between the planets themselves. This work, done before the development of computers, was an extremely laborious task. A good fraction of one's lifetime work could be spent on a single calculation.



**Figure 33** *Planetary orbit when the gravitational force is modified to a*  $1/r^{1.9}$  *force.* 

The orbit of the planet Mercury provided a good test of these calculations because its orbital ellipse is more extended than that of the other close-in planets. The more extended an ellipse, the easier it is to observe a precession. (You cannot even detect a precession for a circular orbit.) Mercury's orbit has a small but observable precession. Its orbit precesses by an angle that is slightly less than .2 degrees every *century*. This is a very small precession which you could never detect in one orbit. But the orbit of Mercury has been observed for about 3000 years, or 30 centuries. That is over a 5 degree precession which is easily detectable.

When measuring small angles, astronomers divide the degree into 60 *minutes of arc*, and for even smaller angles, divide the minute into 60 *seconds of arc*. One second of arc, 1/3600 of a degree, is a very small angle. A basketball 30 miles distant subtends an angle of about 1 second of arc. In these units, Mercury's orbit precesses about 650 seconds of arc per century.

By 1900, astronomers doing Newtonian mechanics calculations could account for all but 43 seconds of arc per century precession of Mercury's orbit as being caused by the influence of neighboring planets. The 43 seconds of arc discrepancy could not be explained. One of the important predictions of Einstein's relativistic theory of gravity is that it predicts a 43 second of arc per century precession of Mercury's orbit, a precession caused by a change in the gravitational force law and not due to neighboring planets. Einstein used this explanation of the 43 seconds of arc discrepancy as the main experimental foundation for his relativistic theory of gravity when he just presented it in 1915. The importance of the Eddington eclipse expedition in 1917 is that a completely new phenomena, predicted by Einstein's theory, was detected.

(The Eddington expedition verified more than just the fact that light is deflected by the gravitational attraction of a star. You can easily construct a theory where the energy in the light beam is related to mass via the formula  $E = mc^2$ , and then use Newtonian gravity to predict a deflection. Einstein's General Relativity predicts a deflection twice as large as this modified Newtonian approach. The Eddington expedition observed the larger prediction of General Relativity, providing convincing evidence that General Relativity rather than Newtonian gravity was the more correct theory of gravity.)

#### Exercise 18

Start with your Orbit 1 program, modify the exponent in the gravitational force law, and see what happens. Begin with a small modification so that you can see how to plot the results. (If you make a larger modification, you will have to change the plotting window to get interesting results.)

(To get the 43 seconds of arc per century precession of Mercury's orbit, using a modified gravitational force law, the force should be proportional to  $1/r^{2.00000016}$  instead of  $1/r^2$ .)

# CONSERVATION OF ANGULAR MOMENTUM

With the ability to work with realistic satellite orbits rather than just the circular orbits, we will be able to make significant tests of the laws of conservation of angular momentum and of energy, as applied to satellite motion. In this section, we will first see how Kepler's second law of planetary motion is a direct consequence of the conservation of angular momentum, and then do some calculations with the Orbit 1 program to see that a satellite's angular momentum is in fact conserved—does not change as the satellite goes around the earth. In the next section we will first take a more general look at the idea of a conservation law, and then apply this discussion to the conservation of energy for satellite orbits.

Recall that Kepler's second law of planetary motion states that a line from the sun to the planet, the radius vector, sweeps out equal areas in equal times. For this to be true when the planet is in an elliptical orbit, the planet must move faster when in close to the sun and the radius vector is short, and slower when far away and the radius vector is long.

To intuitively see that this speeding up and slowing down is a consequence of the conservation of angular momentum, one can modify the three dumbbell experiment we used to demonstrate the conservation of angular momentum. In this demonstration the instructor uses only one dumbbell. After a student assists the instructor in getting his rotation started, the instructor extends the dumbbell out to full arm's reach, for instance, when he is facing the class, and pulls his arm in when he is facing away as shown in Figure (34). Some practice is needed to maintain this pattern and not lose one's balance.

The rather expected result of this demonstration is that the instructor rotates more slowly when his arm is far out, and more rapidly when his arm is in close. If we associate the dumbbell with a satellite orbiting the earth, we see the same speeding up as the lever arm about the axis of rotation is reduced, and slowing down as the lever arm is increased.

A fairly simple geometrical construction demonstrates that the rule about the radius vector sweeping out equal areas in equal times is precisely what is required for conservation of angular momentum. In Figure (35a) we have plotted an elliptical satellite orbit showing the position of the planet for two different equal time intervals. The time intervals  $\Delta t$  are short enough that we can fairly accurately represent the displacement of the satellite by short, straight lines of length  $v_1\Delta t$  in the upper triangle and  $v_2\Delta t$  in the lower triangle. With this approximation we can represent the areas swept out by the radius vector by triangles as shown by the shaded areas in Figure (35a).





Calculating the area swept out by the planet during a short time interval  $\Delta t$ . (b)

**Figure 34** One dumbbell experiment.

Now the area of a triangle is one half the base times the altitude. If you look at the lower triangle in Figure (35a), and take the side  $v_2\Delta t$  as the base, then the distance labeled  $r_{2\perp}$  is the altitude, as seen in the sketch in Figure (34b). Thus the area of the triangle at position 2 is

area swept  
out at position 2  
in a time 
$$\Delta t$$
 =  $\frac{1}{2}$  (base) × (altitude)  
=  $\frac{1}{2}$  (v<sub>2</sub> $\Delta t$ ) × r<sub>2 $\perp$</sub>  (21)

When the satellite is at position 2 in Figure (35a), moving at a velocity  $\vec{v}_2$ , the distance of closest approach if it continued at the same velocity  $\vec{v}_2$  would be the distance  $r_{2\perp}$ . Thus  $r_{2\perp}$  is the "lever arm" for the motion of the satellite at this point in the orbit.

We get a similar formula for the area of the triangle at position 1. Using Kepler's second law which says that these areas should be equal for equal times  $\Delta t$ , we get

$$\frac{1}{2} (\mathbf{v}_1 \Delta t) \mathbf{r}_{1\perp} = \frac{1}{2} (\mathbf{v}_2 \Delta t) \mathbf{r}_{2\perp}$$
(22)

Dividing Equation 22 through by  $\Delta t$  and multiplying both sides by 2m, where m is the mass of the satellite, gives

$$\mathbf{m}_1 \mathbf{v}_1 \mathbf{r}_{1\perp} = \mathbf{m}_2 \mathbf{v}_2 \mathbf{r}_{2\perp} \tag{23}$$

Recall that the definition of a particle's angular momentum about some axis is the linear momentum  $\vec{p} = m\vec{v}$  times the lever arm  $r_{\perp}$  (see Equations 7–15, 16). Thus the left side of Equation 23 is the satellite's angular momentum at position 1, the right side at position 2. The statement that the satellite sweeps out equal areas in equal times is thus equivalent to the statement that the satellite's angular momentum  $mvr_{\perp}$ has the same value all around the orbit. Like the dumbbell in Figure (34), the satellite moves faster when  $r_{\perp}$  is small, and slower then  $r_{\perp}$  is large, in order to conserve angular momentum. As a direct check of the conservation of angular momentum in the satellite orbit program, note that if a particle is located a distance x from an axis of rotation and is moving in the y direction with a velocity  $v_y$  as shown in Figure (36a), the lever arm about the origin is simply x, and the particle's angular momentum about the origin  $l_a$  is

$$\ell_{a} = mxv_{y} \qquad \begin{array}{c} particle's \ angular \\ momentum in \ Figure (36a) \end{array}$$
(24)

Using the right hand convention illustrated in Figure (7-14), we see that this particle has angular momentum directed up, out of the paper. We will call this positive angular momentum. (You can think of m as a small piece of the bicycle wheel shown in Figure 7-14.)

Now consider a particle of mass m located a distance y from the origin traveling in the -x direction as shown in Figure (36b). By the right hand convention the angular momentum is still positive (you could think of this m as another part of the same bicycle wheel), but the x velocity is now negative. Thus the formula for this particle's angular momentum is

$$\ell_{\rm b} = -\,{\rm myv}_{\rm x} \tag{25}$$

We have to put in the minus (–) sign to counteract the fact that  $v_x$  is negative but  $\ell_b$  is positive.

It turns out that if a particle is in the xy plane at some arbitrary position  $\vec{R} = (x,y)$ , and has some arbitrary velocity  $\vec{v} = (v_x, v_y)$  in the xy plane, then the formula for the angular momentum  $\ell_0$  of the particle about the origin is



You can see that this general result is just a combination of the two special cases we considered in Figures (36) and Equations 24 and 25. (Equation 26 also comes from the formula  $\vec{l} = m\vec{r} \times \vec{v}$  where  $\vec{r} \times \vec{v}$  is the vector cross product of  $\vec{r}$  and  $\vec{v}$ . We will discuss vector cross products in detail later in Chapter 11. For now Equation 26 is all we need.)

With Equation 26, we can easily test whether angular momentum is in fact conserved in our satellite orbit calculations. By the end of the calculational loop, we have already calculated new values of the satellite's x and y coordinates  $R_x$  and  $R_y$ , and x and y velocity components  $v_x$  and  $v_y$ . Thus to calculate the satellite's angular momentum, all we need is the line

$$LET Lz = M * (Rx*Vy - Ry*Vx)$$
(27)

where we are using the name  $L_z$  because we are observing the z component of the satellite's angular momentum, as indicated in Figure (37).

To check that angular momentum is conserved, we could add a print line at the end of the calculational loop like

IF MOD 
$$(I, 40) = 0$$
 THEN PRINT Rx, Ry, Lz (28)

By printing the values of  $R_x$  and  $R_y$  as well as  $L_z$ , we can see where the satellite is in its orbit as well as the value of the angular momentum at that point.



Add lines (27) and (28) to your Orbit 1 program and check that angular momentum is conserved. Use several different initial conditions so that you can check conservation of angular momentum for different elliptical orbits. (Make sure that  $L_z$  is calculated within the calculational loop so that the latest values of  $R_x$ ,  $R_y$ ,  $V_x$  and  $V_y$  are used for each calculation.) Also, if you set the satellite mass m equal to 1, the values for  $L_z$  will be easier to interpret. (The value of the constant m does not matter since you are simply checking that  $L_z$  is constant during the satellite's orbital motion.)

#### **Exercise 20**

The fact that angular momentum is conserved in Exercise 19 should not be too surprising because you have already checked in earlier exercises that the elliptical orbit obeys Kepler's second law, and as we have just seen, Kepler's second law implies conservation of angular momentum. In this exercise, see if angular momentum is also conserved if we modify the gravitational force law as we did in Exercise 19. Take your program from Exercise 19, the one that prints out the values of the angular momentum, change the exponent of r in Newton's law of gravity, and see if angular momentum is conserved while the ellipse is precessing.



**Figure 37** Angular momentum vector of a rotating wheel.

# CONSERVATION OF ENERGY

In addition to angular momentum, there is another quantity that is conserved during a satellite's orbital motion. In Chapter 10, which is completely devoted to the topic of energy, we will discuss techniques for deriving formulas for various forms of energy. But it is not necessary to be able to derive energy formulas in order to be able to appreciate and use the concept.

The fundamental idea behind the concept of energy is that energy is a conserved quantity. To study the conservation of energy is often a more difficult job than studying the conservation of linear or angular momentum, because there are many forms that energy can take, and not all the forms are easy to recognize. But in certain simple examples like the motion of an earth satellite, there are only two forms of energy we have to deal with, and the conservation of energy is easy to observe.

Unlike linear and angular momentum, energy does not point anywhere. Energy is represented by a number, not a vector. You get a bill from your electric company for the amount of electrical energy you used the previous month. The electric company has a formula, based on the reading of your electric meter, for the amount of electrical energy you used. Because energy is conserved, the power company could not create the energy they sold you out of nothing, they probably got the energy either from a nuclear power plant or by burning fossil fuels. If they got the energy from fossil fuels, that energy originally came from the sun, from the combining of hydrogen nuclei to form helium nuclei. If the electricity came from a nuclear power plant, the energy came from the splitting of large uranium or plutonium nuclei into smaller nuclei. The uranium and plutonium nuclei were formed by getting their energy from a supernova explosion that must have occurred over five billion years ago.

In our discussion of energy in Chapter 10, we will see that there is a close analogy between keeping track of your checkbook balance in a bank and keeping a record of the amount of energy a system has. With a bank balance, there is a convention that if your balance is positive, the bank owes you money, and if the balance is negative, you owe the bank money. A zero balance indicates that neither owes each other anything. If the bank is not worried about your credit, it does not make much difference whether your balance is positive, negative or zero, you can still write checks, make deposits, and go about your normal business.

In the way we deal with energy, what we call the zero of energy does not make much difference either. We can think of a power company borrowing energy from a coal company just as it borrows money from a bank. In this sense the power company can have a negative energy balance just as it has a negative bank balance. The fact that energy is conserved means that the power company cannot create energy out of nothing to repay the debt. The difference between the power company and physical systems like satellites in orbit is that we let power companies pay their energy debt with cash, a physical system can increase its energy balance only by getting energy from somewhere else.

In our accounting scheme for energy, some terms are positive and some are negative. The term called *kinetic energy* is always positive. In most circumstances, kinetic energy is given by the formula  $1/2 \text{ mv}^2$  where m is the mass of the object and v the object's speed. Kinetic energy is positive because neither m or  $1/2 \text{ mv}^2$  can become negative.

To observe conservation of energy for satellite motion, it is necessary to account for two forms of energy. One is kinetic energy  $1/2v^2$ , the other is what is called *gravitational potential energy*. Our formula for gravitational potential energy will be  $-Gm_sm_e/r$  where G is the gravitational constant,  $m_s$  and  $m_e$  the masses of the satellite and earth respectively, and r the separation between them. This formula looks much like the gravitational force formula, except that it is proportional to 1/r rather than  $1/r^2$ . What is often upsetting to students when they first encounter the gravitational potential energy formula is the minus sign. How can energy be negative? This is essentially a result of our accounting procedure. The important feature of energy is that it is conserved. If the gravitational potential energy in some part of an orbit becomes more negative, then the kinetic energy has to become more positive so that the total is conserved, i.e., stays constant. As far as energy conservation is concerned, it does not make any difference what the total energy is, as long as it is constant.

At this point we have made no effort to explain where the formulas  $1/2 \text{ mv}^2$  for kinetic energy and  $-\text{Gm}_{s}\text{m}_{e}/\text{r}$  for gravitational potential energy came from. That is a subject for Chapter 10. What we are concerned with now is to see if the *Total Energy*, the sum of these two, is conserved as the satellite moves around its orbit.

$$E_{tot} = \frac{1}{2} mv^{2} - \frac{G m_{s}m_{e}}{r}$$
(29)

We will check for conservation of energy in much the same way we checked for conservation of angular momentum using our Orbit 1 program. Near the end of the calculational loop, after we have calculated the latest values of the satellite position  $\vec{r}$  and velocity  $\vec{v}$ , and have also calculated the corresponding magnitudes r and v, we can add the line

$$LET Etot = Ms*V*V/2 - G*Ms*Me/R \qquad (30)$$

Then we can add a print line like

IF MOD(I, 
$$40$$
) = 0 THEN PRINT Rx, Ry, Etot

By looking at the printed values of  $E_{tot}$  we can see whether this formula for  $E_{tot}$  is conserved as the satellite moves around.

#### Exercise 21

Using the steps described above, check that the satellite's total energy  $E_{tot}$  is conserved. (You will notice slight variations in the value of  $E_{tot}$ , the values are not as steady as they were in the printout of angular momentum. Exercise 22 suggests a way of improving the energy calculation and getting better results.)

As a variation, print out the values of the kinetic energy, potential energy and  $E_{tot}$ . You will see big changes in the kinetic and potential energy, while the sum  $E_{tot}$  remains nearly constant. Start the satellite with different initial conditions and check for energy conservation for different elliptical orbits.

#### Exercise 22

We can obtain a more accurate calculation of the satellite's total energy by slightly modifying the value of v used in the kinetic energy formula. When we put the calculation of  $E_{tot}$  at the end of the calculational loop, we are using the value of v at the end of the time step dt. It turns out that we get a more accurate energy calculation if we use a value of v that is the average of the value we had when we entered the calculational loop and the value a time dt later when we left. This averaging is easily accomplished using the following commands inserted into your calculational loop.

LETVold = V	new line saving old value of v			
LET Vx =				
LET Vy =	your old lines calculating the next new value of v			
LET V = SQR (Vx * Vx + Vy * Vy)				
LET Vnew = V	saving the new value of V			
LETV = (Vold + Vnew)/2	setting V to the average value			
LET Etot = $(Ms*V*V)/2 - G*Ms*Me/R$				

The steps above using the average of  $V_{new}$  and  $V_{old}$  for V in the calculation of the kinetic energy represents the kind of specialized computer trick we have tried to avoid in this text. However, the trick works so well, the improvement in the value of the total energy is so great that it is worth the effort. This is particularly true for project work where a check for conservation of energy is the main check of the validity of the calculation. (You can usually spot computer errors by printing out the total energy, because computer errors almost never conserve energy.

#### Exercise 23 (optional, more like a project)

It turns out that if we modify the formula for the gravitational force, for example changing the exponent of r from + 2 to - 1.9, we also have to modify the formula for the gravitational potential energy in order to observe energy conservation. You will learn in Chapter 10 that the formula for the gravitational potential energy is the integral of the magnitude of the force. We can, for example, obtain our formula for gravitational potential energy from the gravitational force formula by the following integration

$$\int_{-\infty}^{1} \frac{Gm_sm_e}{r^2} dr = -\frac{Gm_sm_e}{r}$$
(31)

If you modify the gravitational force formula, you can do the same kind of integration to get the corresponding potential energy formula. (In Chapter 10 we will have a lot more to say about this integration. For now you can treat the integration as a convenient device for obtaining the potential energy formula. Since the important feature of energy is that it is conserved, if you find from running your Orbit 1 program that the total energy turns out to be conserved, you know you have the correct potential energy formula no matter how it was derived.)

For this exercise, start by modifying the gravitational force law by changing the exponent of r from +2to-1.9. Then run your Orbit 1 program using the formula  $-Gm_sm_e/r$  for potential energy to see that this formula does not work. (Use the accurate version of the program from Exercise 22 so that you can be more confident of the results.)

Then integrate  $Gm_sm_e/r^{1.9}$  to find a new potential energy formula. See if energy is conserved with your new formula. Once this is successful, try some other modification.

# Chapter 9 Applications of Newton's Second Law

In the last chapter our focus was on the motion of planets and satellites, the study of which historically lead to the discovery of Newton's law of motion and gravity. In this chapter we will discuss various applications of Newton's laws as applied to objects we encounter here on earth in our daily lives. This chapter contains many of the examples and exercises that are more traditionally associated with an introductory physics course.

# ADDITION OF FORCES

The main new concepts discussed in this chapter are how to deal with a situation in which several forces are acting at the same time on an object. We had a clue for how to deal with this situation in our discussion of projectile motion with air resistance, where in Figure (1) reproduced here, we saw that the acceleration  $\vec{a}$  of the Styrofoam projectile was the *vector sum* of the acceleration  $\vec{g}$  produced by gravity and the acceleration  $\vec{a}_{air}$  produced by the air resistance

$$\vec{a} = \vec{g} + \vec{a}_{air} \tag{1}$$

If we multiply Equation 1 through by m, the mass of the ball, we get

$$m\vec{a} = m\vec{g} + m\vec{a}_{air} \tag{2}$$

We know that  $\vec{mg}$  is the gravitational force acting on the ball, and it seems fairly clear that we should identify  $\vec{ma}_{air}$  as the force  $\vec{F}_{air}$  that the air is exerting on the ball. Thus Equation 2 can be written



Vector addition of accelerations.

In other words the vector  $m\vec{a}$ , the ball's mass times its acceleration, is equal to the vector sum of the forces acting upon it. More formally we can write this statement in the form



Equation 4 forms the basis of this chapter. The basic rule is that, to predict the acceleration of an object, you first identify all the forces acting on the object. You then take the vector sum of these forces, and the result is the object's mass m times its acceleration  $\vec{a}$ .

When we begin to apply Equation 4 in the laboratory, we will be somewhat limited in the number of different forces that we can identify. In fact there is only one force for which we have an explicit and accurate formula, and that is the gravitational force  $m\vec{g}$  that acts on a mass m. Our first step will be to identify other forces such as the force exerted by a stretched spring, so that we can study situations in which more than one force is acting.



**Figure 2** Spring force balanced by the gravitational force.

#### SPRING FORCES

The simplest way to study spring forces is to suspend a spring from one end and hang a mass on the other as shown in Figure (2). If you wait until the mass m has come to rest, the acceleration of the mass is zero and you then know that the vector sum of the forces on m is zero. In this simple case the only forces acting on m are the downward gravitational force  $m\vec{g}$  and the upward spring force  $\vec{F}_s$ . We thus have by Newton's second law

$$\sum_{i} \vec{F}_{i} = \vec{F}_{s} + m\vec{g} = m\vec{a} = 0$$
 (5)

and we immediately get that the magnitude  $F_s$  of the spring force is equal to the magnitude mg of the gravitational force.

As we add more mass to the end of the spring, the spring stretches. The fact that the more we stretch the spring, the more mass it supports, means that the more we stretch the spring the harder it pulls back, the greater  $F_s$  becomes.

To measure the spring force, we started with a spring suspended from a nail and hung 50 gm masses on the end, as shown in Figure (3). With only one 50 gram mass, the length S of the spring, from the nail to the hook on the mass, was 45.4 cm. When we added

50 gm

50 gm

50 gm



another 50 gm mass, the spring stretched to a length of 54.8 centimeters. We added up to five 50 gram masses and plotted the results shown in Figure (4).

Looking at the plot in Figure (4) we see that the points lie along a straight line. This means that the spring force is linearly proportional to the distance the spring has been stretched.

To find the formula for the spring force, we first draw a line through the experimental points and note that the line crosses the zero force axis at a length of 35.9 cm. We will call this distance the unstretched length  $S_0$ .

Thus the distance the spring has been stretched is  $(S - S_o)$ , and the spring force should be linearly proportional to this distance. Writing the spring force formula in the form

$$\mathbf{F}_{s} = \mathbf{k} \left( \mathbf{S} - \mathbf{S}_{o} \right) \tag{6}$$

all we have left is determine the *spring constant* k.



**Figure 4** *Plot of the length of the spring as a function of the force it exerts.* 

The easy way to find the value of k is to solve Equation 6 for k and plug in a numerical value that lies on the straight line we drew through the experimental points. Using the value  $F_s = 200 \text{ gm} \times 980 \text{ cm/sec}$ = 19.6 × 10<sup>4</sup> dynes when the spring is stretched to a distance S = 73.7 cm gives

$$k = \frac{F_s}{(S - S_o)} = \frac{19.6 \times 10^4 \text{ dynes}}{(73.7 - 35.9) \text{ cm}}$$
$$= 5.18 \times 10^3 \frac{\text{dynes}}{\text{cm}}$$

Equation 6, the statement that the force exerted by a spring is linearly proportional to the distance the spring is stretched, is known as *Hooke's law*. Hooke was a contemporary of Isaac Newton, and was one of the first to suspect that gravitational forces decreased as  $1/r^2$ . There was a dispute between Hooke and Newton as to who understood this relationship first. It may be more of a consolation award that the empirical spring force "law" was named after Hooke, while Newton gets credit for the basic gravitational force law.

Hooke's law, by the way, only applies to springs if you do not stretch them too far. If you exceed the "elastic limit", i.e., stretch them so far that they do not return to the original length, you have effectively changed the spring constant k.



Experimental setup.

#### The Spring Pendulum

The spring pendulum experiment is one that nicely demonstrates that an object's acceleration is proportional to the vector sum of the forces acting on it. In this experiment, shown in Figure (5), we attach one end of a spring to a nail, hang a ball on the other end, pull the ball back off to one side, and let go. The ball loops around as seen in the strobe photograph of Figure (6). The orbit of the ball is improved, i.e., made more open and easier to analyze, if we insert a short section of string between the end of the spring and the nail, as indicated in Figure (5).

This experiment does not appear in conventional textbooks because it cannot be analyzed using calculus there is no analytic solution for this motion. But the analysis is quite simple using graphical methods, and a computer can easily predict this motion. The graphical analysis most clearly illustrates the point we want to make with this experiment, namely that the ball's acceleration is proportional to the vector sum of the forces acting on the ball.

In this experiment, there are two forces simultaneously acting on the ball. They are the downward force of gravity  $\vec{F}_g = m\vec{g}$ , and the spring force  $\vec{F}_s$ . The spring force  $\vec{F}_s$  always points back toward the nail from which the spring is suspended, and the magnitude of the



**Figure 6** Strobe photograph of a spring pendulum.

spring force is given by Hooke's law  $F_s = k(S - S_o)$ . Since we can calibrate the spring before the experiment to determine k and  $S_o$ , and since we can measure the distance S from a strobe photograph of the motion, we can determine the spring force at each position of the ball in the photograph.

In Figure (7) we have transferred the information about the positions of the ball from the strobe photograph to graph paper and labeled the first 17 positions of the ball from – 1 to 15. Consider the forces acting on the ball when it is located at the position labeled 0. The spring force  $\vec{F}_s$  points from the ball up to the nail which in this photograph is located at a coordinate (50, 130). The distance S from the hook on the ball to the nail, the distance we have called the stretched length of the spring, is 93.0 cm. You can check this for yourself by marking off the distance from the edge of the ball to the nail on a piece of paper, and then measuring the separation of the marks using the graph paper (as we did back in Figure (1) of Chapter 3). We measure to the edge of the ball and not the center, because that is where the spring ends, and in calibrating the spring we measured the distance S to the end of the spring. (If we measured to the center of the ball, that would introduce

Nail (50,130)



an error of about 1.3 cm, which produces a noticeable error in our results.)

It turns out that we used the same spring in our discussion of Hooke's law as we did for the strobe photograph in Figure (6). Thus the graph in Figure (4) is our calibration curve for the spring. (The length of string added to the spring is included in the unstretched length  $S_o$ ). Using Equation 6 for the spring force, we get

$$F_{s} = k (S - S_{o})$$
  
= 5.18 × 10<sup>3</sup> dynes/cm × (93.0 - 35.9)cm  
= 29.6 × 10<sup>4</sup> dynes (8)

The direction of the spring force is from the ball to the nail. Using a scale in which  $10^4$  dynes = 1 graph paper square, we can draw an arrow on the graph paper to represent this spring force. This arrow, labeled  $\vec{F}_s$ , starts at the center of the ball at position 0, points toward the nail, and has a length of 29.6 graph paper squares.

Throughout the motion, the ball is subject to a gravitational force  $\vec{F}_g$  which points straight down and has a magnitude mg. For the strobe photograph of Figure (6), the mass of the ball was 245 grams, thus the gravitational force has a magnitude

$$F_g = mg = 245 \text{ gm} \times 980 \frac{\text{cm}}{\text{sec}^2}$$
$$= 24.0 \times 10^4 \text{ dynes}$$
(9)

This gravitational force can be represented by a vector labeled  $\vec{mg}$  that starts from the center of the ball at position 0, and goes straight down for a distance of 24 graph paper squares (again using the scale 1 square =  $10^4$  dynes.)

The total force  $\vec{F}_{total}$  is the vector sum of the individual forces  $\vec{F}_s$  and  $\vec{F}_g$ 

$$\vec{F}_{\text{total}} = \vec{F}_{\text{s}} + m\vec{g} \tag{10}$$

The vector addition is done graphically in Figure (8a), giving us the total force acting on the ball when the ball is located at position 0. On the same figure we have repeated the steps discussed above to determine the total force acting on the ball when the ball is up at position 10. Note that there is a significant shift in the total force acting on the ball as it moves around its orbit.

According to Newton's second law, it is this total force  $\vec{F}_{total}$  that produces the ball's acceleration  $\vec{a}$ . Explicitly the vector m $\vec{a}$  should be equal to  $\vec{F}_{total}$ . To check Newton's second law, we can graphically find the ball's acceleration  $\vec{a} \cdot \vec{a}$  at any position in the strobe photograph, and multiply the mass m to get the vector m $\vec{a}$ .

In Figure (8b) we have used the techniques discussed in Chapter 3 to determine the ball's acceleration vector  $\vec{a}\Delta t^2$  at positions 0 and 10. (Recall that for graphical work from a strobe photograph, we had  $\vec{a} = (\vec{s}_2 - \vec{s}_1)/\Delta t^2$  or  $\vec{a}\Delta t^2 = (\vec{s}_2 - \vec{s}_1)$ , where  $\vec{s}_1$  is the previous and  $\vec{s}_2$  the following displacement vec-



**Figure 8a** Force vectors.

tors and  $\Delta t$  the time between images.) At position 0 the vector  $\vec{a} \Delta t^2$  has a length of 5.7 cm as measured directly from the graph paper. Since  $\Delta t = .1$  sec for this strobe photograph, we have, with  $\Delta t^2 = .01$ ,

$$a\Delta t^2 = a \times .01 = 5.7 \text{ cm}$$
  
 $a = \frac{5.7 \text{ cm}}{.01 \text{ sec}^2} = 570 \frac{\text{cm}}{\text{sec}^2}$ 

Using the fact that the mass m of the ball is 245 gm, we find that the length of the vector ma at position 0 is

$$\left| \text{m}\vec{a} \right|_{\text{at position 0}} = 245 \text{ gm} \times 570 \frac{\text{cm}}{\text{sec}^2}$$
$$= 14.0 \times 10^4 \frac{\text{gm cm}}{\text{sec}^2}$$
$$= 14.0 \times 10^4 \text{ dynes}$$

In Figure (8c) we have plotted the vector m  $\vec{a}$  at position 0 using the same scale of one graph paper square equals  $10^4$  dynes. Since m  $\vec{a}$  has a magnitude of  $14.0 \times 10^4$  dynes, we drew an arrow 14.0 squares long. The direction of the arrow is in the same direction





as the direction of the vector  $\vec{a} \Delta t^2$  at position 0 in Figure (8b). In a similar way we have constructed the vector  $\vec{ma}$  at position 10.

As a comparison between theory and experiment, we have drawn both the vectors  $\vec{F}_{total}$  and  $m\vec{a}$  at positions 0 and 10 in Figure (8c). While the agreement is not exact, it is the best we can expect, considering the accuracy with which we can read the strobe photographs. The important result is that the vectors  $\vec{F}_{total}$  and  $m\vec{a}$  can be seen to closely follow each other as the ball moves around the orbit. (In Exercise 1 we ask you to compare  $\vec{F}_{total}$  and  $m\vec{a}$  at a couple of more positions to see these vectors following each other.) (I once showed a figure similar to Figure (8c) to a mathematician, who observed the slight discrepancy between the vectors  $\vec{F}_{total}$  and  $m\vec{a}$  and said, "Gee, it's too bad the experiment didn't work." He did not have much of a feeling for experimental errors in real experiments.)

#### Exercise 1

Using the data for the strobe photograph of Figure (6), as we have been doing above, compare the vectors  $\vec{F}_{total}$  and ma at two more locations of the ball.



Figure 8c Comparing  $\vec{F}_{total}$  and  $\vec{ma}$ .

# Computer Analysis of the Ball Spring Pendulum

It turns out that using the computer you can do quite a good job of predicting the motion of the ball bouncing on the end of the spring. A program for predicting the motion seen in Figure (7) is listed in the appendix of this chapter. Here all we will discuss are the essential features that you will find in the calculational loop of that program.

The main features of any program that predicts the motion of an object are the following lines, written out in English

! Calculational Loop Let  $\vec{R}_{new} = \vec{R}_{old} + \vec{V}_{old} * dt$ Let  $\vec{F}_1 = \dots$  find forces acting on the object Let  $\vec{F}_2 = \dots$ Let  $\vec{F}_{total} = \vec{F}_1 + \vec{F}_2 + \dots$  find the vector sum of the forces Let  $\vec{a} = \vec{F}_{total}/m$  Newton's second law Let  $\vec{V}_{new} = \vec{V}_{old} + \vec{a} * dt$ Loop Until ... ! Repeat calculation (11) To apply this general structure to the spring pendulum problem, we first have to be able to describe the direction of the spring force  $\vec{F}_s$ . This is done using the vector diagram of Figure (9). The vector  $\vec{Z}$  represents the coordinate of the nail from which the spring is suspended,  $\vec{S}$  the displacement from the nail to the ball, and  $\vec{R}$  the coordinate of the ball. From Figure (9) we immediately get the vector equation

$$\vec{Z} + \vec{S} = \vec{R} \tag{12}$$

which we can solve for the spring length  $\vec{S}$ 

$$\vec{S} = \vec{R} - \vec{Z} \tag{13}$$

From Figure (7) we see that the nail is located at the coordinate (50,130) thus

$$\vec{Z} = (50, 130)$$

Throughout the motion of the ball, the spring force points in the  $-\hat{S}$  direction as indicated in Figure (10), thus the formula for the spring force can be written

$$\vec{\mathbf{F}}_{s} = \left(-\hat{\mathbf{S}}\right) \mathbf{k} \left(\mathbf{S} - \mathbf{S}_{o}\right) \tag{14}$$

where  $k(S - S_o)$  is the magnitude of the spring force determined in Figure (4).





Vector diagram.  $\vec{Z}$  is the coordinate of the nail,  $\vec{R}$  the coordinate of the ball, and  $\vec{S}$  the displacement of the ball from the nail.



**Figure 10** Force diagram, showing the two forces acting on the ball. Using Equation 6 for the spring force, the English calculational loop for the spring pendulum becomes

! Calculational loop for spring pendulum

Let 
$$\vec{R}_{new} = \vec{R}_{old} + \vec{V}_{old} * dt$$
  
Let  $\vec{S} = \vec{R} - \vec{Z}$   
Let  $\vec{F}_s = -\hat{S} * k * (S - S_o)$   
Let  $\vec{F}_g = mg$   
Let  $\vec{F}_{total} = \vec{F}_s + \vec{F}_g$   
Let  $\vec{A} = \vec{F}_{total}/m$   
Loop Until . . . (15)

A translation into BASIC of the lines for calculating  $\vec{S}$  and  $\vec{F}_s$  would be, for example,

LET Sx = 
$$Rx - Zx$$
  
LET Sy =  $Ry - Zy$   
LET S = SQR (Sx \* Sx + Sy \* Sy)  
LET Fsx = (-Sx / S) \* k \* (S - So)  
LET Fsy = (-Sy / S) \* k \* (S - So) (16)

x=100

The rest of the program, discussed in the Appendix, is much like our earlier projectile motion programs, with a new calculational loop. In Figure (11) we have plotted the results of the spring pendulum program, where the crosses represent the predicted positions of the ball and the squares are the experimental positions. If you slightly adjust the initial conditions for the motion of the ball, you can make almost all the crosses fall within the squares. How much adjustment of the initial conditions you have to do gives you an indication of the size of the errors involved in determining the positions of the ball from the strobe photograph.

#### Analytic Solution

If you pull the ball straight down and let go, the ball bounces up and down in a periodic motion that can be analyzed using calculus. The resulting motion is called a *sinusoidal oscillation* which we will discuss in considerable detail in Chapter 14. You will see that if you can use calculus to obtain an analytic solution, there are many ways to use the results. The oscillatory spring motion serves as a model for describing many phenomena in physics.



#### Figure 11

Output from the ball spring program. The crosses are the points predicted by the computer program, while the black squares represent the experimental data points. The program in the Appendix illustrates how the data points can be plotted on the same diagram with your computer plot.

# THE INCLINED PLANE

Galileo discovered the formulas for projectile motion by using an inclined plane to slow the motion down, making it easier to measure positions and velocities. He studied rolling balls, whereas we wish to study sliding objects using a frictionless inclined plane.

The frictionless inclined plane was more or less a figment of the imagination of the authors of introductory textbooks, at least until the development of the air track. And even with an air track some small effects of friction can be observed. We will discuss the inclined plane here because it illustrates a useful technique for analyzing the forces on an object, and because it leads to some interesting laboratory experiments.

As a simple experiment, place a book, a floppy disk, or some small object under one end of an air track so that the track is tilted at an angle  $\theta$  as shown in Figure (12). If you keep the angle  $\theta$  small, you can let the air cart bounce against the bumper at the end of the track without damaging anything.

To analyze the motion of the air cart, it helps to exaggerate the angle  $\theta$  in our drawings of the forces involved as we have done this in Figure (13). The first step in handling any Newton's law problem is to identify all the forces involved. In this case there are two forces acting on the air cart; the downward force of gravity mg and the force  $\vec{F}_p$  of the plane against the cart.

The main feature of a frictionless surface is that it can exert only normal forces, i.e., forces perpendicular to the surface. (Any sideways forces are the result of "friction".) Thus  $\vec{F}_p$  is perpendicular to the air track, inclined at an angle  $\theta$  away from the vertical direction.



**Figure 12** *Tilted air cart.* 

What makes the analysis of this problem different from the motion of the spring pendulum discussed in the last section is the fact that the cart is constrained to move along the air track. This tells us immediately that the cart accelerates along the track, and has no acceleration perpendicular the track. If there is no perpendicular acceleration, there must be no net force perpendicular to the track. From this fact alone we can determine the magnitude of the force  $\vec{F}_p$  exerted by the track.

Before we do any calculations, let us set up the problem in such a way that we can take advantage of our knowledge that the cart moves only along the track. Without thinking, we would likely take the x axis to be in the horizontal direction and the y axis in the vertical direction. But with this choice the cart has a component of velocity in both the x and y directions.

The analysis is greatly simplified if we choose one of the coordinate axes to lie along the plane. In Figure (14), we have chosen the x axis to lie along the plane, and decomposed the downward gravitational force into an x component which has a magnitude  $mg \sin\theta$  and a - y component of magnitude  $mg \cos\theta$ .

Now the analysis of the problem is easy. Starting with Newton's law in vector form, we have

$$m\vec{a} = \Sigma \vec{F}_i = m\vec{g} + \vec{F}_p$$
(17)

Separating Equation 17 into its x and y components, we get



$$ma_x = (m\vec{g})_x = mg\sin\theta$$
 (18a)

$$ma_{y} = 0$$
  
=  $F_{p} - (m\vec{g})_{y}$   
=  $F_{p} - mg\cos\theta$  (18b)

where we set  $a_y = 0$  because the cart moves only in the x direction.

From Equation 18b we immediately get

$$F_{p} = mg\cos\theta \tag{19}$$

as the formula for the magnitude of the force the plane exerts on the cart.

Of more interest is the formula for  $a_x$  which we immediately get from Equation 18a

$$a_{\rm x} = g\sin\theta \tag{20}$$

We see that the cart has a constant acceleration down the plane, an acceleration whose magnitude is equal to the acceleration due to gravity, but reduced by a factor  $\sin\theta$ . It is this reduction that slows down the motion, and allowed Galileo to study motion with constant acceleration using the crude timing devices available to him at that time.



Choosing the x axis to lie along the plane.

#### Exercise 2

A one meter long air track is set at an angle of  $\theta$  = .03 radians. (This was done by placing a 3 millimeter thick floppy disk under one end of the track.

(a) From your knowledge of the definition of the radian, explain why, to a high degree of accuracy, the sin  $\theta$  and  $\theta$  are the same for these small angles.

(b) The cart is released from rest at one end of the track. How long will it take to reach the other end. (You can consider this to be a review of the constant acceleration formulas.)



**Portrate of Galileo** 



**Galileo's Inclined plane** 

Above photos from the informative web page http://galileo.imss.firenze.it/museo/b/egalilg.html

# FRICTION

If you do the experiment suggested in Exercise 2, measuring the time it takes the cart to travel down the track when the track is tilted by a very small angle, the results are not likely to come out very close to the prediction. The reason is that for such small angles, the effects of "friction" are noticeable even on an air track.

In introductory physics texts, the word "friction" is used to cover a multitude of sins. With the air track, there is no physical contact between the cart and track. But there are air currents that support the cart and come out around the edge of the cart. These air currents usually slow the air cart down, giving rise to what we might call friction effects.

In common experience, skaters have as nearly a frictionless surface as we are likely to find. The reason that you experience little friction when skating is not because ice itself is that slippery, but because the ice melts under the blade of the skate and the skater travels along on a fine ribbon of water. The ice melts due to the pressure of the skate against the ice. Ice is a peculiar substance in that it expands when it freezes. And conversely, you can melt it by squeezing it. If, however, the temperature is very low, the ice does not melt at reasonable pressures and is therefore no longer slippery. At temperatures of  $40^{\circ}$  F below zero, roads on ice in Alaska are as safe to drive on as paved roads.

When two solid surfaces touch, the friction between them is caused by an interaction between the atoms in the surfaces. In general, this interaction is not understood. Only recently have computer models shed some light on what happens when clean metal surfaces interact. Most surfaces are quite "dirty" at an atomic scale, contaminated by oxides, grit and whatever. It is unlikely that one will develop a comprehensive theory of friction for real surfaces. Friction, however, plays too important a role in our lives to be ignored. Remember the first time you tried to skate and did not have a surface with enough friction to support you. To handle friction, a number of empirical rules have been developed. One of the more useful rules is "if it squeaks, oil it". At a slightly higher level, but not much, are the formulas for friction that appear in introductory physics text books. Our lack of respect for these formulas comes from the experience of trying to verify them in the laboratory. There is some truth to them, but the more accurately one tries to verify them, the worse the results become. With this statement in mind about the friction formulas, we will state them, and provide one example. Hundreds of examples of problems involving friction formulas can be found in other introductory texts.

# **Inclined Plane with Friction**

In our analysis of the air cart on the inclined track, we mentioned that a frictionless surface exerts only a normal force on an object. If there is any sideways force, that is supposed to be a friction force  $\vec{F}_f$ . In Figure (15) we show a cart on an inclined plane, with a friction force  $\vec{F}_f$  included. The normal force  $\vec{F}_n$  is perpendicular to the plane, the friction force  $\vec{F}_f$  is parallel to the plane, and gravity still points down.



**Figure 15** *Friction force acting on the cart.* 

To analyze the motion of the cart when acted on by a friction force, we write Newton's second law in the usual form

$$\vec{ma} = \Sigma \vec{F}_i = \vec{mg} + \vec{F}_n + \vec{F}_f$$
(21)

The only change from Equation 13 is that we have added in the new force  $\vec{F}_f$ .

Since the motion of the cart is still along the plane, it is convenient to take the x axis along the plane as shown in Figure (16). Breaking Equation 21 up into x and y components now gives

$$ma_x = \Sigma F_x = mg \sin \theta - F_f$$
 (22a)

$$ma_y = \Sigma F_y = -mg\cos\theta + F_n = 0$$
 (22b)

From 22b we get,

$$F_n = mg\cos\theta \tag{23}$$

which is the same result as for the frictionless plane.

The new result comes when we look at motion down the plane. Solving 22a for  $a_x$  gives

$$a_{\rm x} = g\sin\theta - F_{\rm f}/m \tag{24}$$

Not surprisingly, the friction force reduces the acceleration down the plane.



#### **Coefficient of Friction**

To go any further than Equation 24, we need some values for the magnitude of the friction force  $F_f$ . It is traditional to assume that  $F_f$  is proportional to the force  $F_n$  between the surfaces. Such a proportionality can be written in the form

$$F_f = \mu F_n \tag{25}$$

where the proportionality constant  $\mu$  is called the *coefficient of friction*.

Equation 25 makes the explicit assumption that the friction force does not depend on the speed at which the object is moving down the plane. But it is easy to show that this is too simple a model. It is harder to start an object sliding than to keep it sliding. This is why you should not jam on the brakes when trying to stop a car suddenly. You should keep the tires rolling so that there is no sliding between the surface of the tire and the surface of the road.

The difference between non slip or static friction and sliding friction is accounted for by saying that there are two different coefficients of friction, the *static* coefficient  $\mu_s$  which applies when the object is not moving, and the *kinetic* coefficient  $\mu_k$  which applies when the objects are sliding. For common surfaces like a rubber tire sliding on a cement road, the static coefficient  $\mu_s$  is greater than the sliding or kinetic coefficient  $\mu_k$ .

Let us substitute Equation 25 into Equation 24 for the motion of an object down an inclined plane, and then see how the hypothesis that  $F_f$  is proportional to  $F_n$  can be tested in the lab. Using Equation 25 and 24 gives

$$a_{x} = g \sin \theta - F_{f}/m$$
$$= g \sin \theta - \mu F_{n}/m$$

Using  $F_n = mg \cos \theta$  gives

$$a_{x} = g \sin \theta - \mu g \cos \theta$$
$$= g \left( \sin \theta - \mu \cos \theta \right)$$
(26)

Equation 26 clearly applies only if  $\sin \theta$  is greater than  $\mu \cos \theta$  because friction cannot pull the object back up the plane.

If we have a block on an inclined plane, and start with the plane at a very small angle, so that  $\sin \theta$  is much less than  $\mu \cos \theta$ , the block will sit there and not slide. If you increase the angle until  $\sin \theta = \mu \cos \theta$ , with  $\mu$ the static coefficient of friction, the block should just start to slide. Thus  $\mu_s$  is determined by the condition

 $\sin \theta = \mu_s \cos \theta$ 

or dividing through by  $\cos \theta$ 

$$\mu_{\rm s} = \tan \theta_{\rm s} \tag{27}$$

where  $\theta_s$  is the angle at which slipping starts.

After the block starts sliding,  $\mu$  is supposed to revert to the smaller coefficient  $\mu_k$  and the acceleration down the plane should be

$$a_{x} = g \sin \theta - \mu_{k} g \cos \theta \qquad (26a)$$

Supposedly one can then determine the kinetic coefficient  $\mu_k$  by measuring the acceleration  $a_x$  and using Equation 26a for  $\mu_k$ .

If you try this experiment in the lab, you may encounter various difficulties. If you try to slide a block down a reasonably smooth board, you may get fairly consistent results and obtain values for  $\mu_s$  and  $\mu_k$ . But if you try to improve the experiment by cleaning and smoothing the surfaces, the results may become inconsistent because clean surfaces have a tendency to stick rather than slide.

The idea that friction forces can be described by two coefficients  $\mu_s$  and  $\mu_k$  allows the authors of introductory physics texts to construct all kinds of homework problems involving friction forces. While these problems may be good mental exercises, comparable to solving challenging crossword puzzles, they are not particularly appropriate for an introductory physics course. The reason is that the formula  $F_f = \mu F_n$  is an over simplification of a complex phenomena. A decent treatment of friction effects belongs in a more advanced engineering oriented course where there is time to study the limitations and applicability of such a rule.

Another favorite device of the authors of introductory texts is the massless string (or rope). The idea that a string has a small mass compared to the object to which it is attached is usually a very good approximation. And strings and ropes are convenient devices for transferring a force from one object to another.

In addition, strings have the advantage that you can immediately tell the direction of the force they transmit. The force has to be along the direction of the string or rope, for a string cannot pull sideways. We used this idea when we discussed the motion of a golf ball swinging in a circle on the end of a string. The string could only pull in along the direction of the string toward the center of the circle. From this we concluded that the force acting on the ball was also toward the center of the circle, in the direction the ball was accelerating.

To see how to analyze the forces transmitted by strings and ropes, consider the example of two children pulling on a rope in a game of tug of war show in Figure (17). Let the child labeled 1 be pulling on the rope with a force  $\vec{F}_1$  and child labeled 2 pulling with a force  $\vec{F}_2$ . Assuming that the rope is pulled straight between them, the forces  $\vec{F}_1$  and  $\vec{F}_2$  will be oppositely directed.

Applying Newton's second law to the rope, and assuming that the force of gravity on the rope is much smaller than either  $\vec{F}_1$  or  $\vec{F}_2$  and therefore can be neglected, we have

If we now assume that the rope is effectively massless, we get

$$\vec{F}_1 + \vec{F}_2 = 0 \tag{28}$$

Thus  $\vec{F}_1$  and  $\vec{F}_2$  are equal in magnitude and oppositely directed. (Note that if there were a net force on a massless rope, the rope would have an infinite acceleration.)

A convenient way to analyze the effects of a taut rope or string is to say that there is a tension T in the rope, and that this tension transmits the force along the rope. In Figure (18) we have redrawn the tug of war and included the tension T. The point where child 1 is holding the rope is subject to the left directed force  $\vec{F}_1$ exerted by the child and the right directed force caused by the tension T in the rope. The total force on this point of contact is  $\vec{F}_1 + \vec{T}_1$ . Since the point of contact is massless, we must have  $\vec{F}_1 + \vec{T}_1 = 0$  and therefore the tension T on the left side of the rope is equal to the magnitude of  $\vec{F}_1$ . A similar argument shows that the tension force  $\vec{T}_2$  exerted on the second child is equal to the magnitude of  $\vec{F}_2$ . And since the magnitude of  $\vec{F}_1$ and  $\vec{F}_2$  are equal, the tension forces must also be equal.

Isaac Newton noted that when a force was transmitted via a massless medium, like our massless rope, or the force of gravity, the objects exerted equal and opposite forces (here  $\vec{T}_1$  and  $\vec{T}_2$ ) on each other. He called this the *Third Law of Motion*. We will have more to say about Newton's third law in our discussion of systems of particles in Chapter 11.)



 $m_{\text{rope}} \vec{a}_{\text{rope}} = \left( \vec{F}_1 + \vec{F}_2 \right)$ 

**Figure 17** *Tug of war*.



**Figure 18** *Tension T in the rope.* 

#### THE ATWOOD'S MACHINE

As an example of using a string to transmit forces, consider the device shown in Figure (19) which is called an *Atwood's Machine*. It simply consists of two masses at the ends of a string, where the string runs over a pulley. We will assume that the pulley is massless and the bearings in the pulley frictionless so that the only effect of the pulley is to change the direction of the string.

To predict the motion of the objects in Figure (19) we start by analyzing the forces on the two masses. Both masses are subject to the downward force of gravity,  $m_1 \vec{g}$  and  $m_2 \vec{g}$  respectively. Let the tension in the string be T. As a result of this tension, the string exerts an upward force  $\vec{T}$  on both blocks as shown. (We saw in the last section that this force  $\vec{T}$  must be the same on both masses.)

Applying Newton's second law to each of the masses, noting there is only motion in the y direction, we get

$$m_1 a_{1y} = T - m_1 g$$
  
 $m_2 a_{2y} = T - m_2 g$ 
(29)

In Equations 29, we note that there are three unknowns T,  $a_{1y}$  and  $a_{2y}$ , and only two equations. Another relationship is needed. This other relationship is supplied by the observation that the length  $\ell$  of the string, given from Figure (19a) is





An Atwood's machine consists of two masses suspended from a string looped over a pulley. The acceleration is proportional to the difference in mass of the two objects.

$$\ell = \mathbf{h}_1 + \mathbf{h}_2 \tag{30}$$

does not change. Differentiating Equation 30 with respect to time and setting  $d\ell/dt = 0$  gives

$$0 = \frac{d\ell}{dt} = \frac{dh_1}{dt} + \frac{dh_2}{dt} = v_1 + v_2$$

where  $v_1 = dh_1/dt$  is the velocity of mass 1, etc. Differentiating again with respect to time gives

$$0 = \frac{dv_1}{dt} + \frac{dv_2}{dt} = a_1 + a_2$$
(31)

Thus the desired relationship is

$$\mathbf{a}_1 = -\mathbf{a}_2 \tag{32}$$

(You might say that it is obvious that  $a_1 = -a_2$ , otherwise the string would have to stretch. But if you are dealing with more complicated pulley problems, it is particularly convenient to write down a formula for the total length of the string, and differentiate to obtain the needed extra relationship between the accelerations.)

Using Equation 32 in 29 we get

$$m_1 a_{1y} = T - m_1 g$$
  
 $-m_2 a_{1y} = T - m_2 g$  (32b)



Figure 19b Forces involved.

Solving 32a for T to get  $T = m_1(a_{1y} + g)$  and using this in Equation 32b gives

$$-m_2a_{1y} = m_1a_{1y} + m_1g - m_2g$$

or

$$a_{1y} = g\left(\frac{m_1 - m_2}{m_1 + m_2}\right)$$
 (33)

From Equation 33 we see that the acceleration of mass  $m_1$  is uniform, and equal to the acceleration due to gravity, modified by the factor  $\left(m_1-m_2\right)/\left(m_1+m_2\right)$ .

When you solve a new problem, see if you can check it by seeing if the limiting cases make sense. In Equation 30, if we set  $m_2 = 0$ , then  $a_1 = g$  and we have a freely falling mass as expected. If  $m_1 = m_2$ , then the masses balance and  $a_{1y} = 0$  as expected. When a formula checks out in its limiting cases, as this one did, there is a good chance that the result is correct.

The advantage of an of the Atwood's Machine is that by choosing  $m_1$  close to, but not equal to  $m_2$ , you can reduce the acceleration, making the motion easier to observe, just as Galileo did by using inclined planes. If you reduce the acceleration too much by making  $m_1$  too nearly equal to  $m_2$ , you run the risk that even small friction in the bearings of the pulley will dominate the results.



In a slight complication of the Atwood's Machine, we use two pulleys instead of one as shown in Figure (20). We can treat this problem very much like the preceding example except that the length of the string is  $l(h_1 + 2h_2)$  plus some constant length representing the part of the string that goes over the pulleys and the part that goes up to the ceiling. Calculate the accelerations of masses  $m_1$  and  $m_2$ . For what values of  $m_1$  and  $m_2$  is the system balanced?

#### Exercise 4

If you want something a little more challenging than Exercise 3, try analyzing the setup shown in Figure (21), or construct your own setup. For Figure (21), it is enough to set up the four equations with four unknowns.



**Figure 20** *Pulley arrangement for Exercise 3.* 



**Figure 21** *Pulley arrangement for Exercise 4* 

## THE CONICAL PENDULUM

Our final example in this chapter is the conical pendulum. This is one of our favorite examples because it involves a combination of Newton's second law, circular motion, no noticeable friction, and the predictions can be checked using an old boot, shoelace and wristwatch.

For a classroom demonstration of the conical pendulum, we usually suspend a relatively heavy ball on a thin rope, with the other end of the rope attached to the ceiling as shown in Figure (22). The ball is swung in a circle so that the path of the rope forms the surface of a cone as shown. The aim is to predict the period of the ball's circular orbit.

The distances involved and the forces acting on the ball are shown in Figure (23). The ball is subject to only two forces, the downward force of gravity  $m\vec{g}$ , and the tension force  $\vec{T}$  of the string. If the angle that the string makes with the vertical is  $\theta$ , then the force T has an upward component  $T_y = T\cos\theta$  and a component directed radially inward of magnitude  $T_x = T\sin\theta$ . (We are analyzing the motion of the ball at the instant when it is at the left side of its orbit, and choosing the x axis to point in toward the center of the circle at this instant.)

Applying Newton's second law to the motion of the ball, noting that  $a_y = 0$  since the ball is not moving up and down, gives



**Figure 22** *The conical pendulum.* 

$$ma_{x} = T_{x}$$

$$ma_{y} = 0 = T_{y} - mg$$
(34)

The special feature of the conical pendulum is the fact that, because the ball is travelling in a circle, we know that it is accelerating toward the center of the circle with an acceleration of magnitude  $a = v^2/r$ . At the instant shown in Figure (23), the x direction points toward the center of the circle, thus  $a = a_x$  and we have

$$a_x = v^2/r \tag{35}$$

The rest of the problem simply consists of solving Equations 34 and 35 for the speed v of the ball and using that to calculate the time the ball has to go around. The easy way to solve these equations is to write them in the form

$$T_x = ma_x = \frac{mv^2}{r}$$
(36a)

$$T_{y} = mg \tag{36b}$$

Dividing Equation 36a by 36b and using  $T_x/T_y = T\sin\theta/T\cos\theta = \tan\theta$ , we get

$$\frac{T_x}{T_y} = \tan \theta = \frac{mv^2}{mgr} = \frac{v^2}{gr}$$



**Figure 23** Forces acting on the ball.

Next use the fact that  $\tan \theta = r/h$  to get

$$\tan \theta = \frac{r}{h} = \frac{v^2}{gr}$$

$$v = r \sqrt{\frac{g}{h}}$$
(37)

Finally we note that the period is the distance traveled in one circuit,  $2\pi r$ , divided by the speed v of the ball

$$\frac{\text{period}}{\text{of orbit}} = \frac{2\pi r}{v} = \frac{2\pi r}{r\sqrt{g/h}}$$

$$period = 2\pi \sqrt{\frac{h}{g}}$$

$$(38)$$

The prediction of Equation 38 is easily tested, for example, by timing 10 rotations of the ball and dividing the total time by 10. Note that if the angle  $\theta$  is kept small, then the height h of the ball is essentially equal to the length  $\ell$  of the rope, and we get the formula

period 
$$\approx 2\pi \sqrt{\frac{\ell}{g}}$$
 (39)

Equation 39 is the famous formula for the period of what is called the simple pendulum, where the ball swings back and forth rather than in a circle. Equation 39 applies to a *simple pendulum* only if the angle  $\theta$  is kept small. For large angles, Equation 38 is exact for a conical pendulum, but Equation 39 has to be replaced by a much more complicated formula for the simple pendulum. (We will discuss the analysis of the simple pendulum in Chapter 11 on rotations and oscillations.)

Note that the formula for the period of a simple pendulum depends only on the strength g of gravity and the length  $\ell$  of the pendulum, and not on the mass m or the amplitude of the swing. As a result you can construct a clock using the pendulum as a timing device, where the period depends only on how long you make the pendulum.

#### Exercise 5 Conical Pendulum

Construct a pendulum by dangling a shoe or a boot from a shoelace.

(a) Verify that for small angles  $\theta$ , you get the same period if you swing the shoe in a circle to form a conical pendulum, or back and forth to form a simple pendulum.

(b) Time 10 swings of your shoe pendulum and verify Equation 38 or 39. (You can get more accurate results using a smaller, more concentrated mass, so that you can determine the distance  $\ell$  more accurately.) Try several values of the shoe string length  $\ell$  to check that the period is actually proportional to  $\sqrt{\ell}$ .

#### **Exercise 6**

This is what we like to call a clean desk problem. Clear off your desk, leaving only a pencil and a piece of paper. Then starting from Newton's second law, derive the formula for the period of a conical pendulum.

What usually happens when you do such a clean desk problem is that since you just read the material, you think you can easily do the analysis without looking at the text. But if you are human, something will go wrong, you get stuck somewhere, and may become discouraged. If you get stuck, peek at the solution and finish the problem. Then a day or so later clean off your desk again and try to work the problem. Eventually you should be able to work the problem without peeking at the solution, and at that point you know the problem well and remember it for a long time.

When you are learning a new subject like Newton's second law, it is helpful to be fully familiar with at least one worked out example for each main topic. In that way when you encounter that topic again in your work, in a lecture, or on an exam, you can draw on that example to remember what the law is and how it is applied.

At various points in this course, we will encounter problems that serve as excellent examples of a topic in the course. The conical pendulum is a good example because it combines Newton's second law with the formula for the acceleration of a particle moving in a circle; the prediction can easily be tested by experiment, and the result is the famous law for the period of a pendulum. When we encounter similarly useful examples during the course, they will also be presented as clean desk problems.

# APPENDIX THE BALL SPRING PROGRAM

! ----- Plotting window (x axis = 1.5 times y axis)1 SET WINDOW -40,140,-10,110 ! ----- Draw & label axes BOX LINES 0,100,0,100 PLOT TEXT, AT -3.0 : "0" PLOT TEXT, AT -13,96: "y=100" PLOT TEXT, AT 101,0 : "x=100" ! ----- Experimental constants LET m = 245LET g = 980LET K = 5130LET So = 35.9LET Zx = 50LET Zy = 130! ----- Initial conditions LET Rx = 88.2LET Ry = 42.8LET Vx = (80.2 - 91.1)/(2\*.1)LET Vy = (24.4 - 63.1)/(2\*.1)LET T = 0CALL CROSS ! ----- Computer Time Step LET dt = .001LET i = 0! ----- Calculational loop DO LET  $Rx = Rx + Vx^*dt$ LET  $Ry = Ry + Vy^*dt$ LET Sx = Rx - ZxLET Sy = Ry - ZyLET S = Sqr(Sx\*Sx + Sy\*Sy) Let  $Fs = K^*(S - So)$ LET Fx = -Fs\*Sx/SLET Fy = -Fs\*Sy/S - m\*gLET Ax = Fx/mLET Ay = Fy/m LET  $Vx = Vx + Ax^*dt$ LET Vy = Vy + Ay\*dtLET T = T + dtLET i = i+1IF MOD(i,100) = 0 THEN CALL CROSS PLOT Rx,Ry

LOOP UNTIL T > 1.6

! ----- Plot data DO READ Rx,Ry CALL BOX LOOP UNTIL END DATA DATA 88.2, 42.8 DATA 80.2, 24.4 DATA 68.0, 12.0 DATA 52.9, 8.6 DATA 37.4, 14.7 DATA 24.0, 28.8 DATA 14.2, 47.5 DATA 9.0, 67.0 DATA 8.2, 83.9 DATA 11.1, 95.0 DATA 16.7, 98.8 DATA 23.9, 94.1 DATA 32.2, 81.5 DATA 41.9, 62.1 DATA 52.1, 39.9 DATA 62.2, 19.4 ! ----- Subroutine "CROSS" draws ! a cross at Rx,Ry. SUB CROSS PLOT LINES: Rx-2,Ry; Rx+2,Ry PLOT LINES: Rx,Ry-2; Rx,Ry+2 END SUB ! ----- Subroutine "BOX" draws ! a cross at Rx,Ry. SUB BOX PLOT LINES: Rx-1,Ry+1; Rx+1,Ry+1 PLOT LINES: Rx-1, Ry-1; Rx+1, Ry-1 PLOT LINES: Rx-1,Ry+1; Rx-1,Ry-1 PLOT LINES: Rx+1,Ry+1; Rx+1,Ry-1 END SUB

END

The new feature is the READ statement at the top of this column. Each READ statement reads in the next values of Rx and Ry from the DATA lines below. We then call BOX which plots a box centered at Rx,Ry. The LOOP statement has this plotting continue until we run out of data. (In Figure 11, we filled in the boxes with a paint program to make them stand out.)

# Chapter 10 Energy

In principle, Newton's laws relating force and acceleration can be used to solve any problem in mechanics involving particles whose size ranges from that of specks of dust to that of planets. In practice, many mechanics problems are too difficult to solve if we try to follow all the details and analyze all the forces involved. For instance  $\vec{f} = m\vec{a}$  presumably applies to the motion of the objects involved in the collision of two automobiles, but it would be an enormous task to study the details of the collision by analyzing all the forces involved.

In a complicated problem, we cannot follow the motion of all the individual particles; instead we look for general principles that follow from Newton's laws and apply these principles to the system of particles as a whole. We have already discussed two such general principles: the laws of conservation of linear and angular momentum. We have found that if two cars traveling on frictionless ice collide and stick together, we can use the law of conservation of linear momentum to calculate their resulting motion. We do not have to know how they hit or any other details of the collision.

In our discussion of satellite motion, we saw that there was another quantity, which we called energy, that was conserved. Our formula for the total energy of the satellite was  $E_{total} = 1/2 mv^2 - Gmm_e/r$  where  $1/2 mv^2$  was called the **kinetic energy** and

 $-Gmm_e/r$  the gravitational potential energy of the satellite. We saw that  $E_{total}$  did not change its value as the satellite went around its orbit.

It turns out that energy is a much more complex subject than we might suspect from the discussion of satellite motion. There are many forms of energy, such as electrical energy, heat energy, light energy, nuclear energy and various forms of potential energy. Sometimes there is a simple formula for a particular form of energy, but sometime it may be hard even to figure out where the energy has gone. Despite the complexity, one simple fact remains, if we look hard enough we find that energy is conserved.

If, in fact, it were not for the conservation of energy, we would not have invented the concept in the first place. Energy is a useful concept only because it is conserved.

What we are going to do in this chapter is first take a more general look at the idea of a conservation law, and then see how we can use energy conservation to develop formulas for the various forms of energy we encounter. We will see, for example, where the formula  $1/2 \text{ mv}^2$  for kinetic energy comes from, and we will show how the formula  $-Gmm_e/r$  for gravitational potential energy reduces to a much simpler formula when applied to objects falling near the surface of the earth.

# CONSERVATION OF ENERGY

Because energy comes in different forms, it is more difficult to state how to compute energy than how to compute linear momentum. But, as we shall see, it is not necessary to state all the formulas for all the different forms of energy. If we know the formula for some forms of energy, we can use the law of conservation of energy to deduce the other formulas as we need them. How a conservation law can be used in this way is illustrated in the following story, told by Richard Feynman in *The Feynman Lectures on Physics* (Vol. I, Addison-Wesley, Reading, Mass., 1963).

"Imagine a child, perhaps 'Dennis the Menace,' who has blocks that are absolutely indestructible, and cannot be divided into pieces. Each is the same as the other. Let us suppose that he has 28 blocks. His mother puts him with his 28 blocks into a room at the beginning of the day. At the end of the day, being curious, she counts the blocks very carefully, and discovers a phenomenal law—no matter what he does with the blocks, there are always 28 remaining! This continues for a number of days, until one day there are only 27 blocks, but a little investigating shows there is one under the rug—she must look everywhere to be sure that the number of blocks has not changed.

One day, however, the number appears to change there are only 26 blocks. Careful investigation indicates that the window was open, and upon looking outside, the other two blocks are found. Another day careful count indicates that there are 30 blocks! This causes considerable consternation, until it is realized that Bruce came to visit, bringing his blocks with him, and he left a few at Dennis' house. After she had disposed of the extra blocks, she closes the window, does not let Bruce in, and then everything is going along all right, until one time she counts and finds only 25 blocks. However, there is a box in the room, a toy box, and the mother goes to open the toy box, but the boy says, 'No, do not open my toy box,' and screams. Mother is not allowed to open the toy box. Being extremely curious, and somewhat ingenious, she invents a scheme! She knows that a block weighs 3 ounces, so she weighs the box at a time when she sees 28 blocks, and it weighs 16 ounces. The next time she wishes to check, she weighs the box again, subtracts 16 ounces and divides by 3. She discovers the following:

$$\begin{pmatrix} number \ of \\ blocks \ seen \end{pmatrix} + \frac{weight \ of \ box - 16 \ oz}{3 \ oz}$$
$$= constant$$

There then appear to be some gradual deviations, but careful study indicates that the dirty water in the bathtub is changing its level. The child is throwing blocks into the water, and she cannot see them because it is so dirty, but she can find out how many blocks are in the water by adding another term to her formula. Since the original height of the water was 6 inches and each block raises the water a quarter of an inch, this new formula would be

$$\begin{pmatrix} number \ of \\ blocks \ seen \end{pmatrix} + \frac{weight \ of \ box - 16 \ oz}{3 \ oz}$$

$$+ \frac{height \ of \ water - 6 \ inches}{1/4 \ inch}$$
(1)

= constant

In the gradual increase in the complexity of her world, she finds a whole series of terms representing ways of calculating how many blocks are in places where she is not allowed to look. As a result of this, she finds a complex formula, a quantity which **has to be computed**, which always stays the same in her situation." Similarly, we will find a series of terms representing ways of calculating various forms of energy. Unlike the story, where some blocks are actually seen, we cannot see energy; all of the terms in our equation for energy must be computed. But if we have included enough terms and have not neglected any forms of energy, the numerical value of all the terms taken together will not change; that is, we will find that energy is conserved.

It is not necessary, however, to start with the complete energy equation. We will begin with one term. Then, as the complexity of our world increases, we will add more terms to the equation so that energy remains conserved.

## MASS ENERGY

On earth, the greatest supply of useful energy ultimately comes from the sun, mainly as sunlight, which is a form of radiant energy. The energy we obtain from fossil fuel, such>é" coal and wood, and the energy we get from hydroelectric dams came originally from the sun. On a clear day, the sun delivers as much energy to half a square mile of tropical land as was released by the first atomic bomb. In about 1 millionth of a second, the sun radiates out into space an amount of energy equal to that used by all of mankind during an entire year.

The sun emits radiant energy at such an enormous rate that if it burned like a huge lump of coal, it would last about 5000 years before burning out. Yet the sun has been burning at nearly its present rate for over 5 billion years and should continue burning for another 5 billion years. How the sun could emit all of this energy was explained in 1905 when Einstein discovered that mass and energy are related through the well-known equation

$$E = mc^2$$
(2)

where **E** is energy, **m** mass, and **c** the speed of light.

The sun's source of energy is the tiny fraction of its mass that is being converted continually to radiant energy through nuclear reactions. Similar processes occur when the hydrogen bomb is exploded. To indicate the amount of energy that is in principle available as mass energy, imagine that the mass of a 5–cent piece (5 gm) could be converted entirely into electrical energy. This electrical energy would be worth several million dollars. The problem is that we do not have the means available to convert mass completely into a useful form of energy. Even in the nuclear reactions in the sun or in the atomic or hydrogen bombs, only a few tenths of 1% of the mass is converted to energy.

Since most of the energy in the universe is in the form of mass energy, we shall begin to develop our equation for energy with Einstein's formula  $E = mc^2$ . As we mentioned, we will add terms to this equation as we discover formulas for other forms of energy.

# **Ergs and Joules**

Our first step will be to use the Einstein energy formula to obtain the dimensions of energy. In the CGS system of units we have

$$E = m gm \times c^2 \frac{cm^2}{sec^2} = mc^2 \frac{gm cm^2}{sec^2}$$

The set of dimensions  $gm cm^2/sec^2$  is called an *erg*.

$$1 \frac{\mathrm{gm}\,\mathrm{cm}^2}{\mathrm{sec}^2} = 1 \,\mathrm{erg} \quad (CGS \,\,\mathrm{units})$$

In the MKS system of units, we have

$$E = m kg \times c^2 \frac{m^2}{\sec^2} = mc^2 kg \frac{m^2}{\sec^2}$$

where the set of dimensions of kg  $m^2/sec^2$  is called a *joule*.

$$1 \frac{\text{kg m}^2}{\text{sec}^2} = 1 \text{ joule} \quad (MKS \text{ units})$$

It turns out that for many applications the MKS joule is a far more convenient unit of energy than the CGS erg. A 100-watt light bulb uses 100 joules of energy per second, or 1 billion ergs of energy per second. The erg is too small a unit of energy for many applications, and it is primarily for this reason that the MKS system of units is more often used than the CGS system. This is particularly true when dealing with electrical phenomena.

#### Exercise 1

(a) Use dimensions to determine how many ergs there are in a joule. (Check your answer against the statement that a 100-watt bulb uses 100 joules or  $10^9$  ergs of energy per second.)

(b) As you may have guessed, a 1 watt light bulb uses 1 joule of energy per second. How many joules of energy does a 1000 watt bulb or heater use in one hour. (This amount of energy is called a **kilowatt hour** (abbreviated **kwh**) and costs a home owner about 10 cents when supplied by the local power company.)

(c) If a 5-cent piece (which has a mass of 5 grams) could be converted entirely to energy, how many kilowatt hours of energy would it produce? What would be the value of this energy at a rate of 10¢ per kilowatt hour?

# **KINETIC ENERGY**

From the recoil definition of mass (Chapter 6), we saw that the mass of an object increases with speed, becoming very large when the speed of the object approaches the speed of light. The formula for the increase in mass with speed was simply

$$m = \frac{m_0}{\sqrt{1 - v^2/c^2}} \tag{6-14}$$

where  $m_0$  is the mass of the particle at rest (the *rest mass*). When we combine this formula with Einstein's equation  $E = mc^2$ , we get as the equation for the energy of a moving particle

$$E = mc^{2} = \frac{m_{0}c^{2}}{\sqrt{1 - v^{2}/c^{2}}}$$
(3)

According to Equation (3), when a particle is at rest (v = 0), its energy is given by

$$E_0 = m_0 c^2 \quad rest energy \tag{4}$$

This energy  $m_0c^2$  is called the *rest energy* of the particle. As a particle begins to move, its mass, and therefore its energy, increases. The *extra* energy that a particle acquires as a result of its motion is called *kinetic energy*. If mc<sup>2</sup> is the total energy, then the formula for the particle's kinetic energy is

$$\frac{\text{kinetic}}{\text{energy}} = \frac{\text{total}}{\text{energy}} - \frac{\text{rest}}{\text{energy}}$$
$$\text{KE} = \text{mc}^2 - \text{m}_0 \text{c}^2 \tag{5}$$

#### **Example 1**

The muons in the motion picture *Time Dilation of the*  $\mu$ –*Meson (Muon) Lifetime* moved at a speed of .995c. By what factor did their mass increase and what is their kinetic energy?

**Solution**: The first step is to calculate  $\sqrt{1 - v^2/c^2}$  for the muons. An easy way to do this is as follows:

v = .995 c  

$$\frac{v}{c}$$
 = .995 = 1 - .005  
 $\frac{v^2}{c^2}$  = (1 - .005)<sup>2</sup>  
= 1 - 2(.005) + (.005)<sup>2</sup>  
= 1 - .01 + .000025

We have neglected .000025 compared to .01 because it is so much smaller. We now have

$$1 - \frac{v^2}{c^2} \approx 1 - (1 - .01) = .01$$
$$\sqrt{1 - \frac{v^2}{c^2}} \approx \sqrt{.01} = .1 = \frac{1}{10}$$

(This procedure is discussed in more detail in the section on approximation formulas in Chapter 1.)

Now that we have  $\sqrt{1 - v^2/c^2} = 1/10$  for these muons, we can calculate their relativistic mass

$$m = \frac{m_0}{\sqrt{1 - v^2/c^2}} = \frac{m_0}{1/10} = 10 m_0$$

Thus the mass of the muons has increased by a factor of 10. The total energy of the muons is

$$\mathbf{E} = \mathbf{m}\mathbf{c}^2 = (10\mathbf{m}_0)\mathbf{c}^2 = 10(\mathbf{m}_0\mathbf{c}^2)$$

Hence, their total energy is also 10 times their rest energy. Their increase in energy, or their kinetic energy, is

$$KE = mc^{2} - m_{0}c^{2} = 10m_{0}c^{2} - m_{0}c^{2}$$
$$KE = 9m_{0}c^{2}$$

This kinetic energy  $9m_0c^2$  is the amount of *additional energy* that is required to get muons moving at a speed v = 0.995c.

#### **Exercise 2**

Assume that an electron is traveling at a speed v = .99995c.

(a) What is  $\sqrt{1 - v^2/c^2}$  for this electron?

(b) By what factor has its mass increased over its rest mass?

(c) By what factor has its total energy increased over its rest energy?

(d) The rest mass of an electron is  $m_0 = 0.911 \times 10^{-27} \text{ gm}$ . What is its rest energy ( in ergs)?

(e) What is the total energy (in ergs) of this electron?

(f) What is the kinetic energy of this electron in ergs?

#### **Slowly Moving Particles**

In Example 1, where the particle (muon) was moving at nearly the speed of light, we determined its increase in mass and its kinetic energy by calculating  $\sqrt{1 - v^2/c^2}$ . However, when a particle is mov-

ing much slower than the speed of light, for instance, 1000 mi/sec or less, there is an easier way to calculate the energy of the object than by evaluating  $\sqrt{1 - v^2/c^2}$  directly.

In the section on approximation formulas in Chapter 1, it was shown that when v/c is much less than 1, then we can use the approximate formula

$$\frac{1}{\sqrt{1-\alpha}} \approx 1 + \frac{\alpha}{2} \tag{1-25}$$

to get

$$\frac{1}{\sqrt{1 - v^2/c^2}} \approx 1 + \frac{v^2}{2c^2}$$
(6)

The approximate formula  $(1 + v^2/2c^2)$  is much easier to use than  $1/\sqrt{1 - v^2/c^2}$ . Moreover, if v/c is a small number, then the formula is quite accurate, as illustrated in Table 1. It should be noted however that when v becomes larger than about .1c, the approximation becomes less accurate. When we reach v = c, the exact formula is  $1/\sqrt{1 - v^2/c^2} = \infty$  but the approximate formula gives  $1 + v^2/2c^2 = 1.5$ . At this point the approximate formula is no good at all!

v	value of exact formula $\frac{1}{\sqrt{1 - v^2/c^2}}$	value of approximate formula $1 + \frac{v^2}{2c^2}$
.01c	1.000050003	1.000050000
.1c	1.005037	1.005000
.2c	1.0206	1.0200
.3c	1.048	1.045
.5c	1.148	1.125
.7c	1.41	1.25
.9c	2.30	1.40
.99c	7.1	1.49
c	$\infty$	1.5

Table 1	Numerical check	of the Approximation Formula	1	≈ 1+-	$v^2$
	Numeriour oncon		$\sqrt{1 - v^2/c^2}$	··· 1 · · · · · · · · · · · · · · · · ·	$2c^2$

If we use Equation (6), the total energy of a particle becomes

$$E = mc^{2}$$

$$= m_{0}c^{2} \left(\frac{1}{\sqrt{1 - v^{2}/c^{2}}}\right) \quad exact formula$$

$$\approx m_{0}c^{2} \left(1 + \frac{v^{2}}{2c^{2}}\right) \qquad approximate formula$$

$$\approx m_{0}c^{2} + m_{0}e^{2}\frac{v^{2}}{2e^{2}}$$

The factor  $c^2$  cancels in the second term, and we are left with the approximate formula

$$E \approx m_0 c^2 + \frac{1}{2} m_0 v^2 \qquad \begin{array}{c} approximate formula\\ for particles moving\\ at speeds less than\\ about .1c \end{array}$$
(7)

Since Equation (7) contains the approximation made in Equation (6), it is not valid for particles traveling faster than about one tenth of the speed of light. For particles traveling at nearly the speed of light, we must use  $E = m_0 c^2 / \sqrt{1 - v^2 / c^2}$ . But for particles traveling as slowly as a few thousand miles an hour or less, Equation (6) is so accurate that any error would be difficult to detect.

For all but the last section of this chapter, we will confine our discussion to the energy of objects traveling at slow speeds, where Equation (7) is not only accurate, but is the simplest equation to use. When we look at this equation, we can see that the mass energy  $E = mc^2$  is now written in two distinct parts  $m_0c^2$ , which is the rest mass energy, and  $1/2m_0v^2$ , which is the energy of motion or kinetic energy

$$\mathbf{E} = \mathbf{m}_0 \mathbf{c}^2 \left( \begin{array}{c} rest\\ energy \end{array} \right) + \frac{1}{2} \mathbf{m}_0 \mathbf{v}^2 \left( \begin{array}{c} kinetic\\ energy \end{array} \right)$$
(7a)

Written in this way, our equation for total energy is beginning to resemble Equation (1), which was used to determine the number of blocks in Dennis' room. We now have two terms representing two different kinds of energy.

10-7

It is worth noting that, at one time, only the kinetic energy term  $1/2m_0v^2$  in Equation 7 was recognized as a form of energy. Before 1905, it was not known that  $m_0c^2$  should be included in the equation for conservation of energy, because no one had ever observed the rest mass of an object to change. The first evidence that the rest energy had to be included came from the study of nuclear reactions. In these reactions enormous amounts of energy were released, producing a detectable change in the nuclear rest masses.

So long as an object is moving at a speed of .1c or less, the kinetic energy of that object will be far less than its rest mass energy. For example, let us compare the kinetic energy to the rest mass energy of a 10-gm pistol bullet that travels with a speed of about 300 m/sec. Using MKS units, we find that the bullet's kinetic energy (KE) is

KE = 
$$\frac{1}{2}m_0 v^2$$
  
=  $\frac{1}{2} \times .01 \text{ kg} \times (300 \text{ m/sec})^2$   
= 450 joules

This is enough to allow a bullet to penetrate a plank. The rest mass energy  $E_0$  of the bullet is

$$E_0 = m_0 c^2$$
  
= .01 kg  $(3 \times 10^8 \text{ m/sec})^2$   
= 9 × 10<sup>14</sup> joules

This is the amount of energy released in a moderatesized atomic bomb.

#### **Exercise 3**

For the preceding example of a 10 gram bullet:

- a) at 10 cents per kilowatt hour, what is the value of the bullet's kinetic energy?
- b) what is the value of its rest energy?

c) how fast would the bullet be traveling if it had twice as much kinetic energy?



#### Figure 1

Falling Weight. When a weight is dropped it gains kinetic energy. This kinetic energy comes from the energy we stored in the object when we lifted it up to a height h.

# GRAVITATIONAL POTENTIAL ENERGY

Let us continue our search for terms to add to our equation for energy. Suppose that a ball of mass m is dropped from a height h above the floor, as shown in Figure (1). Immediately before the ball hits the floor, it has a rest energy  $m_0c^2$ , and a kinetic energy  $1/2m_0v^2$ . Immediately before the ball was dropped, however, it had the same rest energy  $m_0c^2$  but no kinetic energy. Where did the kinetic energy that it possessed just before it hit the floor come from?

If we were observant, we might have noted that some effort was needed to lift the ball from the floor to a height h. As the ball was lifted a new kind of energy was being stored. This new form of energy, which was released when the ball was dropped, is called *gravitational potential energy*. When it is included, our equation for energy becomes

$$E_{total} = m_0 c^2 + \frac{1}{2} m_0 v^2 + \frac{gravitational}{potential}$$
(8)

To find the formula for the gravitational potential energy, we will assume that energy is conserved and that the total energy of the ball, immediately before it is released, is equal to the total energy of the ball immediately before it hits the ground.

When a ball is dropped from a height h, it accelerates downward with a constant acceleration g until it hits the floor. Thus we can use the constant acceleration formulas (see Appendix 1 in Chapter 4.)

$$s = v_i t + \frac{1}{2}at^2$$

$$v_f = v_i + at$$

with  $v_i = 0$ , a = g, and s = h we get

$$h = \frac{1}{2}gt^2 \tag{12}$$

$$v_{\rm f} = gt \tag{13}$$
Substituting  $t = v_f / g$  from Equation 13 into Equation (12) gives

$$h = \frac{1}{2}g\left(\frac{v_{f}^{2}}{g^{2}}\right) = \frac{v_{f}^{2}}{2g}$$
$$\frac{1}{2}v_{f}^{2} = gh \qquad (14)$$

Multiplying Equation 14 through by  $m_0$  gives

$$\frac{1}{2}m_0 v_f^2 = m_0 gh$$
 (15)

Suppose that we use  $m_0$ gh as the formula for gravitational potential energy. (The greater h, the higher we have lifted the ball, the more potential energy we have stored in it.)

$$m_0 gh = \begin{cases} formula for & near the \\ gravitational & surface of \\ potential energy & the earth \end{cases} (16)$$

Before the ball is released, its total energy is in the form of rest energy and gravitational potential energy

$$E_{\text{total}} \begin{pmatrix} before \\ release \end{pmatrix} = m_0 c^2 + m_0 gh$$
(17)

Just before the ball hits the floor, where it has kinetic energy but no potential energy (since h = 0), the total energy is

$$E_{\text{total}} \begin{pmatrix} just \ before \\ hitting \ floor \end{pmatrix} = m_0 c^2 + \frac{1}{2} m_0 v_f^2 \quad (18)$$

At first, Equations 17 and 18 for total energy look different; but since  $1/2m_0v_f^2 = m_0gh$  (Equation 15), they give the same numerical value for the ball's total energy. Thus, we conclude that we have chosen the correct formula for calculating gravitational potential energy.

#### **Exercise 4**

Call  $v_2$  the speed of the ball when it has fallen halfway to the floor.

(a) Explain why the ball's total energy, when it has fallen halfway to the floor, is

$$\mathbf{E}_{\text{total}} \begin{pmatrix} halfway\\ down \end{pmatrix} = \mathbf{m}_0 \mathbf{c}^2 + \frac{1}{2} \mathbf{m}_0 \mathbf{v}_2^2 + \mathbf{m}_0 \mathbf{g} \left(\frac{\mathbf{h}}{2}\right)$$

(b) Calculate  $v_2$  (just as we calculated  $v_f$ ) and show that the total energy of the ball when halfway down is the same as when it was released, or just before it hit the floor.

#### Exercise 5

Show that the formula for gravitational potential energy has the dimensions of joules (in the MKS system) and ergs (in the CGS system).

#### **Exercise 6**

What is the gravitational potential energy (in joules and ergs) of a 100–gm ball at a height of 2 meters above the floor? (Measure h starting from the floor.)

What happens to the energy after the ball has hit the floor and is lying at rest? At this point, it no longer has kinetic energy or gravitational potential energy. Now what should we add to our equation to maintain conservation of energy? In this case, we have to look "under the rug," in the "dirty water," and "out the window" all at once. When the ball hit the floor, we heard a thump; thus, some of the ball's energy has been dissipated as sound energy. We find that there is a dent in the floor; hence we know that some of the energy has gone into rearranging the molecules in that part of the floor. Also, because the bottom of the ball and the floor underneath became slightly warmer after the ball hit the floor, we conclude that some of the energy was converted into heat energy. (In some collisions, such as when a mining pick strikes a stone, we see what looks like a spark, which shows us that some of the kinetic energy has been changed into radiant energy, or light.)

After the ball hits the floor, the formula for total energy becomes as complicated as

$$E_{\text{total}} = m_0 c^2 + \frac{1}{2} m_0 v^2 + m_0 gh$$
  
+ sound energy  
+ energy to cause a dent (19)

+ heat energy + light energy

Because energy can appear in so many forms that are often difficult to detect, it was not until many years after Newton that conservation of energy was established as a general law. The law of conservation of energy is used to solve only those problems where very little energy "escapes" in a form that is difficult to detect. In a complicated collision problem we can calculate only how much energy is "lost," that is, changed to other forms of energy.

On an atomic scale, however, we do not have to think of energy as being "lost" because the various forms of energy are more easily detected. For example, we will see in Chapter 17 that the heat energy and sound energy are primarily the kinetic energy of the atoms and molecules; thus, these do not appear as separate forms of energy. It is on this small scale that the law of conservation of energy may be most accurately verified.

On the other hand, if we can neglect the effects of friction and air resistance, the law of conservation of energy can be used to solve mechanics problems that would otherwise be difficult to solve. We will illustrate this with two examples in which gravitational potential energy  $m_0$ gh is converted into kinetic energy  $1/2m_0v^2$  and vice versa.

#### Notation

Since our discussion for the remainder of this chapter will deal with objects moving at speeds much less than the speed of light, objects whose mass m is very nearly equal to the rest mass  $m_0$ , we will stop writing the subscript 0 for the rest mass. With this notation, our formulas for kinetic energy and gravitational potential energy are simply  $1/2 \text{ mv}^2$  and mgh. Only when we discuss objects like atomic particles whose speeds can become relativistic, will we be careful to distinguish the rest mass  $m_0$  from the total mass m.

## Example 2

Consider a simple pendulum consisting of a ball swinging on the end of a string, as shown in Figure (2). When the ball is released from a height h it has a potential energy  $m_0$ gh. As the ball swings down toward the bottom, h decreases and the ball loses potential energy but gains kinetic energy. At the bottom the original potential energy mgh has been entirely converted into kinetic energy 1/2mv<sup>2</sup>. Then the ball climbs again, gaining potential energy but losing kinetic energy.





Application of conservation of energy to pendulum motion. The speed at B can be found by equating the kinetic energy at B  $(1/2mv^2)$  to the potential energy lost in going from A to B (mgh).

Finally, at position C, the ball has swung back up to a height h and all the kinetic energy has been changed to potential energy. The ball stops momentarily at position C, and the swing is reversed. Eventually, however, the pivot becomes warm and air currents are set up by the swinging pendulum; thus, the pendulum itself gradually loses energy and finally comes to rest.

As long as we can neglect air resistance and friction in the pivot we can use the conservation of energy equation to calculate the speed of the ball at position B. Before the ball is released

$$E_{totalA} = m_0 c^2 + mgh$$

At position B, where h = 0

$$E_{\text{total B}} = m_0 c^2 + \frac{1}{2} m v_{\text{B}}^2$$

If energy is conserved

$$E_{\text{total A}} = E_{\text{total B}}$$
$$m_0 c^2 + mgh = m_0 c^2 + \frac{1}{2} mv^2$$

Note that since  $m_0c^2$  did not change, it does not enter into this calculation. Here we could apply the conservation of energy equation without considering the rest energy. We now have

$$mgh = \frac{1}{2}mv_{B}^{2}$$
$$v_{B}^{2} = 2gh$$
$$v_{B} = \sqrt{2gh}$$

## Example 3

It should be noted that we are able to calculate the speed of the ball in the preceding example without an analysis of the forces involved. An even more striking example of conservation of energy that would be nearly impossible to analyze in terms of forces is that of a skier traveling down a very icy hill. If he is not an experienced skier, he may not know how to dissipate some of his kinetic energy as heat and sound by scraping the edges of his skis against the ice. If he is not able to dissipate energy, then no matter how he turns, no matter how twisted a path he takes, when he reaches bottom, all his potential energy mogh will have been converted to kinetic energy  $1/2m_0v^2$ , in which case his speed at the bottom of the hill will be  $\sqrt{2}$ gh . To see why an inexperienced skier should not try icy hills, consider that if the hill has a 500-ft rise, his speed at the bottom will be 179 ft/sec or 122 mi/hr. This result is computed not from the details of the skier's path, but from the knowledge that he was not able to dissipate energy. As we mentioned at the beginning of the chapter, the conservation of energy is one of the general principles of mechanics that can be applied successfully without knowing all the details involved in the physical situation.

#### Exercise 7

A car coasts along a road that leads from the top of a 300–ft–high hill, down through a valley, and up over a 200 ft high hill. Assume that the car does not dissipate energy through friction and air resistance.

(a) If the car starts at rest from atop the higher hill, how fast will it be traveling when it reaches the top of the lower hill  $(g = 32 \text{ ft/sec}^2)$ ?

(b) If the car is initially moving at 80 ft/sec (55 mi/hr) when it starts coasting at the top of the higher hill, how fast will the car be moving when it reaches the top of the lower hill?

# WORK

Let us take another look at the example where we dropped a ball of mass m from a height h above the floor as shown in Figure (3). At the height h, the ball had a gravitational potential energy mgh. Just before hitting the floor, all this gravitational potential energy had been converted to kinetic energy  $1/2 \text{ mv}^2$ . We know that the ball speeded up, accelerated, because gravity was exerting a downward force mg on the ball as it fell.

There appears to be a coincidence in this example. Gravity pulls down on the ball with a force of magnitude mg, the ball falls a distance h, and the ball gains a kinetic energy equal to  $(mg) \times h$ . In this example the energy that gravity supplies to the ball by pulling down on it is equal to the gravitational force (mg) times the distance h over which the force acted. Is this a coincidence, or does this example provide a clue as to the way in which forces supply energy?

In this case, where we have a constant force mg, and the ball moves in the direction of the force for a distance h, the increase in energy is the force times the distance.

In more general examples, however, the situation can be more complex. If the object is not moving in the direction of the force, then only the component of the force in the direction of motion adds energy to the object. And if the force is not constant, we have to break the problem into many small steps, and calculate the energy gained in each step. We shall see that calculus provides powerful techniques to handle these situations.



**Figure 3** *A ball, subject to a gravitational force mg, falling a distance h, gains a kinetic energy mgh.* 

We will begin the discussion with the introduction of a new term which we will call *work*. In some ways this is an unfortunate choice of a word, for everyone has their own idea of what "work" is, and it seldom coincides with the physicist's definition. In the physicist's definition, *a force does work on an object when it adds energy to the object*. More explicitly, the work a force does is equal to the energy that the force supplies. In the case of the falling ball the gravitational force supplied an amount of energy mgh, therefore that is the work that the gravitational force did as the ball fell.

work done by the force of gravity as the ball fell 
$$\langle 20 \rangle$$

From Equation (20), we see that for the case where we have a constant force, and the object moves in the direction of the force, the work done is equal to the magnitude of the force times the distance moved.



## Exercise 8

Show that force times distance has the same dimensions as energy. (Get the dimensions of energy from  $E = mc^2$ .)

As the first complication, or correction to our definition of work, suppose that the force is not in the same direction as the motion. Suppose, for example, a hockey puck slides for a distance S along frictionless ice as shown in Figure (4). During this motion a gravitational force mg is acting and the puck moves a distance S. But the puck coasts along at constant speed; it does not gain any energy at all. In this case the gravitational force does no work.



Figure 4 The force of gravity does no work on the sliding hockey puck.

The problem with the hockey puck example is that the gravitational force is down and the motion is sideways. In this case the -y directed gravitational force has no component along the x directed motion of the puck. In order for the puck to gain energy, it must accelerate in the x direction, but there is no x component of force to produce that acceleration.

Now let us consider an example where the force is acting opposite to the direction of motion. If we throw a ball up in the air, the ball starts out with the kinetic energy  $1/2 \text{ mv}_0^2$  that we gave it. As the ball rises, gravity acts against the motion of the ball and removes kinetic energy. When the ball has risen to a height h given by mgh =  $1/2 \text{ mv}_0^2$ , all the kinetic energy is gone and the ball stops. The ball has reached the top of the trajectory. This example tells us that when the force is directed opposite to the direction of motion, the work is negative—the force removes rather than adds energy.

## The Dot Product

This is where our discussion has lead so far. We have a quantity called "work" which is a form of energy. It is the energy supplied by a force acting on a moving object. Now energy, given by formulas like  $E = mc^2$ , is a scalar quantity; it is a number that does not point anywhere. But our formula for *work* = *force times distance* involves two vectors, the force  $\vec{F}$  and the distance  $\vec{S}$ . What mathematical way can we combine the two vectors  $\vec{F}$  and  $\vec{S}$  to get a number for the work W? One possibility, that we discussed back in the chapter on vectors, is the scalar or dot product.

$$W = \vec{F} \cdot \vec{S} = |\vec{F}| |\vec{S}| \cos \theta \xrightarrow{\theta} \vec{F}$$
(22)

Mathematically the dot product turns the vectors  $\vec{F}$ and  $\vec{S}$  into a scalar number W. Let us see if  $W = \vec{F} \cdot \vec{S}$ is the correct formula for work. If  $\vec{F}$  and  $\vec{S}$  are in the same direction,  $\theta = 0^\circ$ ,  $\cos \theta = 1$ , and we get

$$W = \vec{F} \cdot \vec{S} = \left| \vec{F} \right| \left| \vec{S} \right| \cos \theta = \left| \vec{F} \right| \left| \vec{S} \right|$$

Applied to the case of a falling ball,  $|\vec{F}| = mg$ ,  $|\vec{S}| = h$ and we get W = mgh which is correct. When we throw the ball up, the angle between the downward force and upward motion is  $\theta = 180^{\circ}$ ,  $\cos \theta = -1$ , and we get

$$W = \vec{F} \cdot \vec{S} = |\vec{F}| |\vec{S}| \cos \theta$$
$$= mgh(-1) = -mgh$$

We now predict that gravity is taking energy from the ball, which is also correct.

Finally, in the case of a hockey puck, the angle  $\theta$  between the -y directed force and the x directed motion is 90°. We have  $\cos \theta = 0$ , so that  $\vec{F} \cdot \vec{S} = 0$  and the gravitational force does no work. Again the formula  $W = \vec{F} \cdot \vec{S}$  works.

#### **Exercise 9**

A frictionless plane is inclined at an angle  $\theta$  as shown in Figure (5). A hockey puck initially at a height h above the ground, slides down the plane. When the puck gets to the bottom, it has moved a distance S = h/cos $\theta$  as shown. (This comes from h = S cos $\theta$ )

a) Verify the formula  $S = h/\cos\theta$  for the two cases  $\theta = 0$  and  $\theta = 90^{\circ}$ . I.e., what are the values for  $h/\cos\theta$  for these two cases, and are the answers correct?

b) Show that the work  $W = \vec{F}_g \cdot \vec{S}$ , done by the gravitational force as the puck slides down the plane, is mgh no matter what the angle  $\theta$  is.

c) Explain the result of part (b) from the point of view of conservation of energy.



**Figure 5** *Diagram for Exercise 9.* 

## **Work and Potential Energy**

In the discussion of energy, physicists tend to use a lot of words like *work, potential energy, kinetic energy*, etc. What we are doing is building a conceptual picture to help us organize a number of physical phenomena and related mathematical equations. You will find that when you see this picture, are familiar with the "jargon", these concepts become easy to use and powerful in their applications. Much of this chapter is to introduce the jargon and develop the picture.

The ideas of work and potential energy are closely related and play critical roles in the picture of energy. Let us discuss some examples simply from the point of view of getting used to the jargon.

Suppose I pick a ball of mass m off the floor and slowly lift it up to a height h. While lifting the ball, I have to just barely overcome the downward gravitational force mg. Therefore I exert an upward directed force of magnitude mg, and I do this for a distance h. Since my upward force and the upward displacement are in the same direction, the work I do, call it  $W_{me}$ , is my force mg times the distance h, or  $W_{me} =$  mgh. Using the ideas of potential energy discussed earlier, we can say that all the energy  $W_{me} =$  mgh that I supplied lifting the ball went into gravitational potential energy mgh.

While I was lifting the ball, gravity was pulling down. The downward gravitational force and the upward displacement were in opposite directions and therefore the work done by the gravitational force was negative. While we are storing gravitational potential energy, gravity does negative work. When we let go of the ball, gravity releases potential energy by doing positive work. Let us consider another example where we store potential energy by doing work against a force. Suppose I tie one end of a spring to a post and pull on the other end as shown in Figure (6). As I stretch the spring, I am exerting a force  $\vec{F}_{me}$  and moving the end of the spring in the same direction. Therefore I am doing positive work on the spring, and this energy is stored in what we can call the "elastic potential energy" of the stretched spring. (We know that a stretched spring has some form of potential energy, for a stretched spring can be used to launch a ball up into the air.)

# **Non-Constant Forces**

Our example above, of storing energy in a spring by stretching it, introduces a new complication. We cannot calculate the work I do  $W_{me}$  in stretching the spring by writing  $W_{me} = \vec{F}_{me} \cdot \vec{S}$ . The problem is that, the farther I stretch the spring, the harder it pulls back (Hooke's law). If I slowly pull the spring out, I have to apply an increasingly stronger force. If we try to use the formula  $W_{me} = \vec{F}_{me} \cdot \vec{S}$ , the problem is what value of  $\vec{F}_{me}$  to use. Do we use the weak  $\vec{F}_{me}$  at the beginning of the pull, the strong one at the end, or some average value.

We could use an average value, but there is a more general way to calculate the work I do. Suppose I wish to pull the spring from an initial position  $x_i$  to a final position  $x_f$ . Imagine that I break this span from  $x_i$  to  $x_f$  into a bunch of small intervals of width  $\Delta x$ , ending at points labeled  $x_0$ ,  $x_1$ , ...  $x_n$  as shown in Figure (7). During each small interval the spring force does not change by much, and I can stretch the spring through that interval by exerting a force equal to the strength of



**Figure 6** *Doing work on a spring.* 

the spring force at the end of the interval. For example in stretching the spring from position  $x_0$  to  $x_1$ , I apply a force of magnitude  $F_s(x_1)$  for a distance  $\Delta x$  and therefore do an amount of work

$$(\Delta W_{me})_1 = F_s(x_1)\Delta x$$

To get out to position  $x_2$ , I increase my force to  $F_s(x_2)$ and apply that force over another interval  $\Delta x$  to do an amount of work

$$(\Delta W_{me})_2 = F_s(x_2)\Delta x$$

If I keep repeating this process until I reach the final position  $x_f$ , the total amount of work I have done is

$$(W_{me})_{total} = (\Delta W_{me})_1 + (\Delta W_{me})_2 + \dots + (\Delta W_{me})_n$$
$$= F_s(x_1)\Delta x + F_s(x_2)\Delta x + \dots + F_s(x_n)\Delta x$$
$$= \sum_{i=1}^n F_s(x_i)\Delta x \qquad (23)$$

In Equation 23, we still have an approximate calculation as long as the intervals  $\Delta x$  are of finite size. We get an exact calculation of the work I do if we take the limit as  $\Delta x$  goes to zero, and the number of intervals goes to infinity. In that limit, the right side of Equation



$$(W_{me})_{total} = -\int_{x_i}^{x_f} F_s(x) dx \qquad (24)$$

The statement of the work we did, Equation 24, can be written more formally by noting that the spring force  $\vec{F}_s(x)$  is actually a vector which points opposite to the direction I pulled the spring. In addition, we should think of each  $\Delta x$  or dx as a small vector displacement  $\vec{\Delta x}$  or  $\vec{dx}$  in the direction I pulled. Since my force was directed opposite to  $\vec{F}_s$ , the work I did during each interval  $\vec{dx}$  can be written as the dot product

$$dW_{me} = \vec{F}_{me} \cdot \vec{dx} = -\vec{F}_{s} \cdot \vec{dx}$$

and the formula for the total work I did becomes

$$(W_{me})_{total} = -\int_{x_i}^{x_f} \vec{F}_s \cdot \vec{dx}$$
(25)

Equation 25 is more general but a bit clumsier to use than 24. To use Equation 25, we would first note that I was pulling along the x axis, and thus dx = dx. Then I would note that the spring force was opposite to the direction I was pulling, so that  $-\vec{F}_s(x) \cdot d\vec{x} = +F_s(x)dx$ 



Figure 7

I can stretch the spring through a series of small intervals of length  $\Delta x$ . In each interval I apply a constant force that is just strong enough to get the spring to the end of the interval.

where  $F_s(x)$  is the formula for the strength of the spring force. That gets me back to Equation (24) and the problem of evaluating the definite integral.

## **Potential Energy Stored in a Spring**

Springs are useful in physics demonstrations and problems because of the simple force law (Hooke's law) which is quite accurately obeyed by real springs. In our study of the motion of a ball on the end of a spring in Chapter 9, we saw that the formula for the strength of the spring force was

$$F_s = K(S - S_0)$$
 (9-6)

where S is the length of the spring and  $S_0$  the unstretched length (the length at which  $F_s$  goes to zero in Figure 9–4).

We can simplify the spring force formula, get rid of the  $S_0$ , by considering a situation where an object is held in an equilibrium position by spring forces. Suppose for example we have a cart on an air track with springs





The spring force  $\vec{F}_{s}$  is always opposite to the displacement x. If the spring is displaced right,  $\vec{F}_{s}$  points left, and vice versa.

connecting the cart to each end of the track as shown in Figure (8). Mark the center of the cart with an arrow, and choose a coordinate system where x = 0 is at the equilibrium position as shown in Figure (8a).

With this setup, the spring force is always a restoring force that is pushing the cart back to the equilibrium position x = 0. If we give the cart a positive displacement as in Figure (8b), we get a left directed or negative spring force. A negative displacement shown in (8c) produces a right directed or positive spring force. And to a high degree of accuracy, the strength of the spring force is proportional to the magnitude of the displacement from equilibrium.

All of these results can be described by the formula

$$F_{s}(x) = -Kx \tag{26}$$

where the minus sign tells us that a positive displacement x produces a negative directed force and vice versa. There is no S<sub>0</sub> or x<sub>0</sub> in Equation 26 because we chose x=0 to be the equilibrium position where  $F_s = 0$ . Equation 26 is what one usually finds as a statement of Hooke's law, and K is called the spring constant.

Equation 26 allows us to easily calculate the potential energy stored in the springs. If I start with the cart at rest at the equilibrium position as shown in Figure (8a), and pull the cart to the right a distance  $x_f$ , the work I do is

$$W_{me} = \int_{x=0}^{x=x_{f}} F_{me} dx = \int_{x=0}^{x=x_{f}} (-F_{s}) dx$$
$$= \int_{x=0}^{x=x_{f}} Kx dx$$
(27)

where I have to exert a force  $F_{me} = -F_s$  to stretch the spring.

In Equation 27, the constant K can come outside the integral, we are left with the integral of xdx which is  $x^2/2$ , and we get

$$W_{me} = K \int_{x=0}^{x=x_{f}} x \, dx = K \frac{x^{2}}{2} \Big|_{0}^{x_{f}} = K \frac{x_{f}^{2}}{2}$$

Noting that all the work I do is stored as "*elastic* potential energy of the spring", we get the formula

Spring potential energy = 
$$K \frac{x^2}{2}$$
 (28)

In Equation 28, we replaced  $x_f$  by x since the formula applies to any displacement  $x_f$  I choose.

#### Exercise 10

If you pull the cart of Figure (8) back a distance  $x_f$  from the equilibrium position and let go, all the potential energy you stored in the cart will be converted to kinetic energy when the cart crosses the equilibrium position x = 0. Use this example of conservation of energy to calculate the speed v of the cart when it crosses x = 0. (Assume that you release the cart from rest.)

Exercise 10, which you should have done by now, illustrates one of the main reasons for bothering to calculate potential energy. It is much easier to predict the speed of the ball using energy conservation than it is using Newton's second law. We can immediately find the speed of the ball by equating the kinetic energy at x = 0 to the potential energy at  $x = x_f$  where we released the cart. To make the same prediction using Newton's second law, we would have to solve a differential equation and do a lot more calculation.

## Exercise 11

With a little bit of cleverness, we can use energy conservation to predict the speed of the cart at any point along the air track. Suppose you release the cart from rest at a distance  $x_f$ , and want to know the cart's speed at, say,  $x_f/2$ . First calculate how much potential energy the cart loses in going from  $x_f$  to  $x_f/2$ , and then equate that to the kinetic energy  $1/2mv^2$  that the cart has gained at  $x_f/2$ .

## **WORK ENERGY THEOREM**

The reason that it is easier to apply energy conservation than Newton's second law is that when we have a formula for potential energy, we have already done much of the calculation. We can illustrate this by deriving what is called the "Work Energy Theorem" where we use Newton's second law to derive a relation between work and kinetic energy. We will first derive the theorem for one dimensional motion, and then see that it is easily extended to motion in three dimensions.

Suppose a particle is moving along the x axis as shown in Figure (9). Let a force  $F_x(x)$  be acting on the particle. Then by Newton's second law

$$F_{x}(x) = ma_{x}(x) = m\frac{dv_{x}(x)}{dt}$$
(29)

Multiplying by dx and integrating to calculate the work done by the force  $F_x$ , we get

$$\int_{i}^{f} F_{x}(x)dx = m \int_{i}^{T} \frac{dv_{x}(x)}{dt}dx$$
(30)

In Equation (30), we are integrating from some initial position  $x_i$  where the object has a speed  $v_{xi}$ , to a position  $x_f$  where the speed is  $v_{xf}$ .



**Figure 9** An x directed force acting on a particle moving in the x direction.

The next step is a standard calculus trick that you may or may not remember. We will first move things around a bit in the integral on the right side of Equation 30:

$$m\int_{i}^{f} \frac{dv_{x}}{dt} dx = m\int_{i}^{f} dv_{x} \left(\frac{dx}{dt}\right)$$
(31)

Next note that  $dx/dt = v_x$ , the x component of the velocity of the particle. Thus the integral becomes

$$m\int_{i}^{f} dv_{x} \left(\frac{dx}{dt}\right) = m\int_{V_{i}}^{V_{f}} v_{x} dv_{x}$$
(32)

After this transformation, we can do the integral because everything is now expressed in terms of the one variable  $v_x$ . Using the fact that the integral of  $v_x dv_x$  is  $v_x^2/2$ , we get

$$\begin{split} m \int_{V_{i}}^{V_{f}} v_{x} dv_{x} &= m \frac{v_{x}^{2}}{2} \Big|_{V_{i}}^{V_{f}} \\ &= \frac{1}{2} m v_{fx}^{2} - \frac{1}{2} m v_{ix}^{2} \end{split}$$
(33)

Using Equations (31) through (33) in Equation (30) gives

$$\begin{bmatrix} x_{f} \\ \int F_{x}(x)dx = \frac{1}{2}mv_{fx}^{2} - \frac{1}{2}mv_{ix}^{2} \\ x_{i} \end{bmatrix}$$
(34)

The left side of Equation 34 is the work done by the force  $F_x$  as the particle moves from position  $x_i$  to position  $x_f$ . The right side is the change in the kinetic energy. Equation 34 tells us that the work done by the force  $F_x$  equals the change in the particle's kinetic energy. This is the basic idea of the work energy theorem.

To derive the three dimensional form of Equation 34, start with Newton's second law in vector form

$$\vec{F} = m\vec{a}$$
 (35)

Take the dot product of Equation 35 with  $\vec{dx}$  and integrate from i to f to get

$$\int_{i}^{f} \vec{F} \cdot \vec{dx} = \int_{i}^{f} m\vec{a} \cdot \vec{dx}$$
(36)

Writing

$$\vec{a} \cdot \vec{dx} = a_x dx + a_y dy + a_z dz$$
(37)

we get

$$\int_{i}^{f} \vec{F} \cdot \vec{dx} = m \int_{i}^{1} \left( \frac{dv_x}{dt} dx + \frac{dv_y}{dt} dy + \frac{dv_z}{dt} dz \right) \quad (38)$$

f

Following the same steps we used to get from Equation 31 to 33, we get

$$\int_{i}^{f} \vec{F} \cdot \vec{dx} = \frac{1}{2} m v_{fx}^{2} - \frac{1}{2} m v_{ix}^{2} + \frac{1}{2} m v_{fy}^{2} - \frac{1}{2} m v_{iy}^{2} + \frac{1}{2} m v_{fz}^{2} - \frac{1}{2} m v_{iz}^{2}$$
(39)

Finally noting that by the Pythagorean theorem

$$v_i^2 = v_{ix}^2 + v_{iy}^2 + v_{iz}^2$$
  

$$v_f^2 = v_{fx}^2 + v_{fy}^2 + v_{fz}^2$$
(40)

we get, using (40) in (39)

$$\int_{i}^{f} \vec{F} \cdot \vec{dx} = \frac{1}{2} m v_{f}^{2} - \frac{1}{2} m v_{i}^{2}$$
(41)

)

which is the three dimensional form of the work energy theorem.

#### **Several Forces**

Suppose several forces  $\vec{F}_1$ ,  $\vec{F}_2$ , ... are acting on the particle as the particle moves from position i to position f. Then the vector  $\vec{F}$  in Equations 35 through 41 is the total force  $\vec{F}_{tot}$  which is the vector sum of the individual forces:

$$\vec{F} = \vec{F}_{tot} = \vec{F}_1 + \vec{F}_2 + \dots$$
 (42)

Our formula for the work done by these forces becomes

$$\int_{i}^{f} \vec{F} \cdot \vec{dx} = \int_{i}^{f} (\vec{F}_{1} + \vec{F}_{2} + ...) \cdot \vec{dx}$$

$$= \int_{i}^{f} \vec{F}_{1} \cdot \vec{dx} + \int_{i}^{f} \vec{F}_{2} \cdot \vec{dx} + \dots \qquad (43)$$

and we see that the work done by several forces is just the numerical sum of the work done by each force acting on the object. Equation 41 now has the interpretation that *the total work done by all the forces acting on a particle is equal to the change in the kinetic energy of the particle*.

## **Conservation of Energy**

To see how the work energy theorem leads to the idea of conservation of energy, suppose we have a particle subject to one force, like the spring force  $\vec{F}_s$  acting on an air cart as shown in Figure (8). If the cart moves from position i to position f, then the work energy theorem, Equation 41 gives

$$\int_{i}^{f} \vec{F} \cdot \vec{dx} = \frac{1}{2} m v_{f}^{2} - \frac{1}{2} m v_{i}^{2}$$
(44)

In our analysis of the spring potential energy, we saw that if I slowly moved the cart from position i to position f, I had to exert a force  $\vec{F}_{me}$  that just overcame the spring force  $\vec{F}_s$ , i.e.,  $\vec{F}_{me} = -\vec{F}_s$ . When I moved the cart slowly, the work I did went into changing the potential energy of the cart. Thus the formula for the change in the cart's potential energy is

$$\left.\begin{array}{l} \text{change in the} \\ \text{potential energy of} \\ \text{the cart when the} \\ \text{cart moves from} \\ \text{position i to position f} \end{array}\right\right\} = \int_{i}^{f} \vec{F} \cdot \vec{dx}$$

$$= -\int_{i}^{f} \vec{F}_{s} \cdot \vec{dx}$$

$$(45)$$

Equation 45 is essentially equivalent to Equation 25 which we derived in our discussion of spring forces.

Spring forces have the property that the energy stored in the spring depends only on the length of the spring, and not on how the spring was stretched. This means that the change in the spring's potential energy does not depend upon whether I moved the cart, or I let go and the spring moves the cart. We should remove  $\vec{F}_{me}$  from Equation 45 and simply express the spring potential energy in terms of the spring force

$$\frac{\text{change in spring}}{\text{potential energy}} = -\int_{i}^{t} \vec{F}_{s} \cdot \vec{dx}$$
 (46)

Equation 46 says that the change in potential energy is **minus** the work done by the force on the object as the object moves from i to f. There is a minus sign because, if the force does positive work, potential energy is released or decreases. We will see that Equation 46 is a fairly general relationship between a force and it's associated potential energy.

We are now ready to convert the work energy theorem into a statement of conservation of energy. Rewrite Equation 44 with the work term on the right hand side and we get

$$0 = \left\{ -\int_{i}^{f} \vec{F}_{s} \cdot \vec{dx} \right\} + \left\{ \frac{1}{2} m v_{f}^{2} - \frac{1}{2} m v_{i}^{2} \right\} \quad (47)$$

The term in the first curly brackets is the change in the particle's potential energy, the second term is the change in the particle's kinetic energy. Equation 47 says that the sum of these two changes is zero

$$0 = \left\{ \begin{array}{c} \text{change in} \\ \text{potential energy} \end{array} \right\} + \left\{ \begin{array}{c} \text{change in} \\ \text{kinetic energy} \end{array} \right\}$$
(47a)

If we define the total energy of the particle as the sum of the particle's potential energy plus its kinetic energy, then the change in the particle's total energy in moving from position i to position f is the sum of the two changes on the right side of Equation 47a. Equation 47a says that this total change is zero, or that the total energy is conserved.

## Conservative and Non-Conservative Forces

We mentioned that the potential energy stored in a spring depends only on the amount the spring is stretched, and not on how it was stretched. This means that the change in potential energy depends only on the initial and final lengths of the spring, and not on how we stretched it. This implies that the integral

$$-\int_{i}^{t} \vec{F}_{s} \cdot \vec{dx}$$

has a unique value that does not depend upon how the particle was moved from i to f.

Gravitational forces have a similar property. If I lift an object from the floor to a height h, the increase in gravitational potential energy is mgh. This is true whether I lift the object straight up, or run around the room five times while lifting it. The formula for the change in gravitational potential energy is

$$\begin{array}{l} \begin{array}{c} \text{change in} \\ \text{gravitational} \\ \text{potential energy} \end{array} \end{array} = -\int_{i}^{f} \vec{F}_{g} \cdot \vec{dx} \\ \\ = -\int_{0}^{h} F_{gy} dy \\ \\ = -\int_{0}^{h} (-mg) dy \\ \\ = mgh \end{array}$$
(48)

Again we have the change in potential energy equal to minus the work done by the force.

Not all forces, however, work like spring and gravitational forces. Suppose I grab an eraser and push it around on the table top for a while. In this case I am overcoming the friction force between the table and the eraser, and we have  $\vec{F}_{me} = -\vec{F}_{friction}$ . The total work done by me as I move the eraser from an initial position i to a final position f is

work I do  
while moving  
the eraser 
$$= -\int_{i}^{f} \vec{F}_{me} \cdot \vec{dx}$$
$$= -\int_{i}^{f} \vec{F}_{friction} \cdot \vec{dx}$$
(49)

There are two problems with this example. The integrals in Equation 49 do depend on the path I take. If I move the eraser around in circles I do a lot more work than if I move it in a straight line between the two points. And when I get to position f, there is no stored potential energy. Instead all the energy that I supplied overcoming friction has probably been dissipated in the form of heat.

Physicists divide all forces in the world into two categories. Those forces like gravity and the spring force, where the integral

$$\int_{i}^{1} \vec{F} \cdot \vec{dx}$$

depends only on the initial and final positions i and f, are called "*conservative*" forces. For these forces there is a potential energy, and the formula for the change in potential energy is minus the work the force does when the particle goes from i to f.

All the other forces, the ones for which the work integral depends upon the path, are called *non-conservative* forces. We cannot use the concept of potential energy for non-conservative forces because the formula for potential energy would not have a unique or meaningful value. The non-conservative forces can do work and change kinetic energy, but as we see in the case of friction, the work ends up as something else like heat rather than potential energy.

It is interesting that on an atomic scale, where energy does not disappear in subtle ways like heat, we almost always deal with conservative forces and can use the concept of potential energy.

# GRAVITATIONAL POTENTIAL ENERGY ON A LARGE SCALE

In our computer analysis of satellite motion, we saw that the quantity  $E_{tot}$ , given by

$$E_{tot} = \frac{1}{2} mv^2 - \frac{GM_em}{r}$$
(50)

was unchanged as the satellite moved around the earth. As shown in Figure (10), m is the mass of the satellite,  $\vec{v}$  its velocity, R its distance from the center of the earth, and  $M_e$  is the mass of the earth. This was our first non trivial example of conservation of energy, where  $1/2 \text{ mv}^2$  is the satellite's kinetic energy, and  $-GM_em/R$  must be the formula for the satellites's gravitational potential energy. Our discussion of the last section suggests that we should be able to obtain this formula for gravitational potential energy by integrating the gravitational force  $|\vec{F}_g| = GM_em/r^2$  from some initial to some final position.

Here on the surface of the earth, the formula for gravitational potential energy is mgh. This simple result arises from the fact that when we lift an object inside a room, the strength of the gravitational force  $\vec{mg}$  acting on it is essentially constant. Thus the work I do lifting a ball a distance h is just the gravitational force mg times the height h. Since this work is stored as potential energy, the formula for gravitational potential energy is simply mgh.

In the case of satellite motion, however, the strength of the gravitational force was not constant. In our first calculation of satellite motion in Chapter 8, the satellite started 1.1 earth radii from the center of the earth and went out as far as r = 5.6 earth radii. Since the gravitational force drops off as  $1/r^2$ , the gravitational force was more than 25 times weaker when the satellite was far away, than when it was launched.



# Zero of Potential Energy

Another difference is that the formula mgh for a ball in the room measures changes in gravitational potential energy starting from the floor where h = 0. In a rather arbitrary way, we have defined the gravitational potential energy to be zero at the floor. This is a convenient choice for people working in this room, but people working upstairs or downstairs would naturally choose their own floors rather than our floor as the zero of gravitational potential energy for objects they were studying.

Since conservation of energy deals only with changes in energy, it does not make any difference where you choose your zero of potential energy. A different choice simply adds a constant to the formula for total energy, and an unchanging or constant amount of energy cannot be detected. The most famous example of this was the fact that a particle's rest energy  $m_0 c^2$  was unknown until Einstein introduced the special theory of relativity, and undetected until we saw changes in rest energy caused by nuclear reactions. In the case of the gravitational potential energy of a ball, if we use the floor downstairs as the zero of gravitational potential energy, we add the constant term (mg)h<sub>floor</sub> to all our formulas for  $E_{tot}$  (where h floor is the distance between floors in this building). This constant term has no detectable effect.

In finding a formula for gravitational potential energy of satellites, planets, stars, etc., we should select a convenient floor or zero of potential energy. For the motion of a satellite around the earth, we could choose gravitational potential energy to be zero at the earth's surface. Then the satellite's potential energy would be positive when its distance r from the center of the earth is greater than the earth radius r<sub>e</sub>, and negative should r become less than re. Such a choice would be reasonable if we were only going to study earth satellites, but the motion of a satellite about the earth is very closely related to the motion of the planets about the sun and the motion of moons about other planets. Choosing  $r = r_e$  as the distance at which gravitational potential energy is zero is neither a general or particularly convenient choice.

In describing the interaction between particles, for example an electron and a proton in a hydrogen atom, the earth and a satellite, the sun and its planets, or the stars in a galaxy, the convenient choice for the zero of potential energy is where the particles are so far apart that they do not interact. If the earth and a rock are a hundred light years apart, there is almost no gravitational force between them, and it is reasonable that they do not have any gravitational potential energy either.

Now suppose that *the earth and the rock are the only things in the universe*. Even at a hundred light years there is still some gravitational attraction, so that the rock will begin to fall toward the earth. As the rock gets closer to the earth it will pick up speed and thus gain kinetic energy. It was the gravitational force of attraction that caused this increase in speed, therefore there must be a conversion of gravitational potential energy into kinetic energy.

This gives rise to a problem. The rock starts with zero gravitational potential energy when it is very far away. As the rock approaches the earth, gravitational potential energy is converted into kinetic energy. How can we convert gravitational potential energy into kinetic energy if we started with zero potential energy?

Keeping track of energy is very much a bookkeeping scheme, like keeping track of the balance in your bank account. Suppose you begin the month with a balance of zero dollars, and start spending money by writing checks. If you have a trusting bank, this works because your bank balance simply becomes negative.

In much the same way, the rock falling toward the earth started with zero gravitational potential energy. As the rock picked up speed falling toward the earth, it gained kinetic energy at the expense of potential energy. Since it started with zero potential energy, and spent some, it must have a negative potential energy balance. From this we see that if we choose gravitational potential energy between two objects to be zero when the objects are very far apart, then the potential energy must be negative when the objects are a smaller distance apart. When we think of energy conservation as a bookkeeping scheme, then the idea of negative potential energy is no worse than the idea of a negative checking account balance. (In the analogy between potential energy and a checking account, the discovery of rest energy  $m_0c^2$  would be like discovering that you had inherited the bank. The checks still work the same way even though your total assets are vastly different.)

Let us now return to Equation (50) and our formula for gravitational potential energy of a satellite

$$\frac{\text{gravitational}}{\text{potential energy}} = -\frac{\text{GM}_{\text{e}}\text{m}}{\text{r}}$$
(50a)

First we see that if the satellite is very far away, that as r goes to infinity, the potential energy goes to zero. Thus this formula does give zero potential energy when the earth and the satellite are so far apart that they no longer interact. In addition, the potential energy is negative, as it must be if the satellite falls in to a distance r, converting potential energy into kinetic energy.

What we have to do is to show that Equation (50a) is in fact the correct formula for gravitational potential energy. We can do that by calculating the work gravity does on the satellite as it falls in from  $r = \infty$  to r = r. This work, which would show up as the kinetic energy of a falling satellite, must be the amount of potential energy spent. Thus the potential energy balance must be the negative of this work. Since the work is the integral of the gravitational force times the distance, we have

gravitational  
potential energy  
at position R 
$$= - \left| \int_{\infty}^{R} \vec{F}_{g} \cdot d\vec{r} \right|$$
  
 $= - \left| \int_{\infty}^{R} \frac{GMm}{r^{2}} dr \right|$  (51)

Equation 51 may look a bit peculiar in the way we have handled the signs. We have argued physically that the gravitational potential energy must be negative, and we know that it must be equal in magnitude to the integral of the gravitational force from  $r = \infty$  to r = R. By noting ahead of time what the sign of the answer must be, we can do the integral easily without keeping track of the various minus signs that are involved. (One minus sign is in the formula for potential energy, another is the dot product since  $\vec{F}_g$  points in and  $d\vec{r}$  out, a third in the integral of  $r^{-2}$ , and more come in the evaluation of the limits. It is not worth the effort to get all these signs right when you know from a simple physical argument that the answer must be negative.) Carrying out the integral in Equation 51 gives

$$\int_{\infty}^{R} \frac{GM_{e}m}{r^{2}} dr = GM_{e}m \int_{\infty}^{R} \frac{dr}{r^{2}}$$
$$= -\frac{GM_{e}m}{r} \Big|_{\infty}^{R} = GM_{e}m \Big(\frac{1}{\infty} - \frac{1}{R}\Big)$$

where we used the fact that the integral of  $1/r^2$  is -1/r. Thus we get

$$\left| \int_{\infty}^{R} \frac{GM_{e}m}{r^{2}} dr \right| = \frac{GM_{e}m}{R}$$

As a result the gravitational potential energy of the satellite a distance R from the center of the earth is - GM<sub>e</sub>m/R as given in Equation 50a.

## Gravitational Potential Energy in a Room

Before we leave our discussion of gravitational potential energy, we should show that the formula –  $GM_em/r$ leads to the formula mgh for the potential energy of a ball in a room. To show this, let us use the formula –  $GM_em/r$  to calculate the increase in gravitational potential energy when I lift a ball from the floor, a distance  $R_e$  from the center of the earth, up to a height h, a distance  $R_e$ + h from the center of the earth, as shown in Figure (11).

We have

$$PE_{at height h} = -\frac{GM_{e}m}{R_{e} + h}$$

$$PE_{at floor} = -\frac{GM_{e}m}{R_{e}}$$

$$Increase_{in PE} = PE_{at h} - PE_{at floor}$$

$$= \left(-\frac{GM_{e}m}{R_{e} + h}\right) - \left(-\frac{GM_{e}m}{R_{e}}\right)$$

$$= GM_{e}m\left(\frac{1}{R_{e}} - \frac{1}{R_{e} + h}\right)$$
(52)

To evaluate the right side of Equation 52, we can write



**Figure 11** A height h above the surface of the earth.

Since  $h/R_e$  is a very small number compared to one, we can use our small number approximation

$$\frac{1}{1+\alpha} \approx 1-\alpha \qquad \text{if } \alpha << 1$$

to write

$$\frac{1}{1 + h/R_e} \approx 1 - \frac{h}{R_e}$$

so that

$$\frac{1}{R_e + h} \approx \frac{1}{R_e} \left( 1 - \frac{h}{R_e} \right) = \frac{1}{R_e} - \frac{h}{R_e^2}$$
(53)

Using Equation 53 in (52) gives

$$\frac{\text{Increase}}{\text{in PE}} \right\} = -GM_{e}m\left[\frac{1}{R_{e}} - \left(\frac{1}{R_{e}} - \frac{h}{R_{e}^{2}}\right)\right]$$
$$= GM_{e}m\left[\frac{h}{R_{e}^{2}}\right]$$
$$= \left(\frac{GM_{e}}{R_{e}^{2}}\right)mh$$

Finally noting that  $GM_e/R_e^2 = g$ , the acceleration due to gravity at the surface of the earth, we get

which is the expected result.

# SATELLITE MOTION AND TOTAL ENERGY

Consider a satellite moving in a circular orbit about the earth, as shown in Figure (12). We want to calculate the kinetic energy, potential energy, and total energy (sum of the kinetic and potential energy) for the satellite. To find the kinetic energy, we analyze its motion, using Newton's laws. The only force acting on the satellite is the gravitational force  $\vec{F_g}$  given by

$$\left|\vec{F_g}\right| = \frac{GMm}{r^2}$$
  $\vec{F_g}$  directed toward the earth

where we now let M = mass of the earth and m = mass of the satellite. Since the satellite is moving at constant speed v in a circle of radius r, its acceleration is  $v^2/r$  toward the center of the circle

$$|\vec{a}| = \frac{v^2}{r}$$
  $\vec{a}$  directed  
toward the earth

Since  $\vec{a}$  and  $\vec{F_g}$  are in the same direction, by Newton's second law

$$\left|\vec{F_g}\right| = m \left|\vec{a}\right|$$
$$\frac{GMm}{r^2} = \frac{mv^2}{r}$$

From this last equation we find that the kinetic energy  $1/2mv^2$  of the satellite is

 $\frac{1}{2}mv^2 = \frac{1}{2}\frac{GMm}{r} \qquad \substack{kinetic\\energy}$ 

The kinetic energy, as always, is positive.



Satellite in a circular orbit.

The gravitational potential energy of the satellite is always negative. Since the satellite is a distance r from the center of the earth, its potential energy is

$$\begin{array}{l} \text{potential} \\ \text{energy} \end{array} = -\frac{\text{GMm}}{\text{r}} \end{array}$$

The total energy of the satellite is

$$E_{\text{total}} = \underset{\text{energy}}{\text{kinetic}} + \underset{\text{energy}}{\text{potential}}$$
$$= \frac{1}{2} \frac{\text{GMm}}{\text{r}} + \left(-\frac{\text{GMm}}{\text{r}}\right)$$
$$E_{\text{total}} = -\frac{1}{2} \frac{\text{GMm}}{\text{r}}$$
(54)

The total energy of a satellite in a circular orbit is negative.

Now consider a satellite in an elliptical orbit. In particular, suppose that the orbit is an extended ellipse, as shown in Figure (13). At apogee, the farthest point from the earth, the satellite is moving very slowly (explain why by using Kepler's law of equal areas). For all practical purposes, the satellite drifts out, stops at apogee, then falls back toward the earth. At apogee, the satellite has *almost no kinetic energy*; at this point its total energy is nearly equal to its negative potential energy

$$E_{total} = -\frac{GMm}{r_{apogee}}$$



#### Figure 13

Satellite in a very eccentric orbit. By Kepler's law of equal areas, a satellite with the above orbit would almost be at rest at apogee. Since the total energy is conserved,  $E_{total}$  remains negative throughout the orbit. If similar satellites are placed in different orbits, the one that goes out the farthest (has the greatest  $r_{apogee}$ ) is the one with the least negative total energy, but all the satellites in elliptical orbits will have a negative total energy.

Suppose an extra powerful rocket is used and a satellite is launched with a positive total energy. In such a case, the positive kinetic energy must always exceed the negative potential energy. No matter how far out the satellite goes, headed for apogee, it will always have some positive kinetic energy to carry it out farther.



#### Figure 14

a) Hyperbolic orbit of comet with positive total energy.

b) Parabolic orbit of comet with zero total energy.

c) Elliptical orbit of comet with slightly negative total energy. (Dashed lines show parabolic orbit for comparison.)

d) Nearly circular orbits of the tightly bound (large negative energy) planets.

Even at enormous distances, where the negative potential energy – GMm/r is about zero, some kinetic energy would still remain, and the satellite would escape from the earth!

By choosing potential energy to be zero when the satellite is very far out, the total energy becomes a meaningful number in itself. If the total energy is negative, the satellite will remain bound to the earth; it does not have sufficient energy to escape. If a satellite launched with positive total energy, it must escape since the negative gravitational potential energy is not sufficiently great to bind the satellite to the earth. If the satellite's total energy is zero, it barely escapes.

The orbits of comets about the sun are interesting examples of orbits of different total energies. It can be shown that when a *satellite's* total energy is positive, its orbit will be in the shape of a *hyperbola*, which is an open-ended curve, as shown in Figure (14a). In this orbit the comet has a positive total energy and never returns.

If the total energy of the comet is zero, the orbit will be in the shape of an open curve, called a*parabola* (Figure 14b). A comet in this kind of orbit will not return either.

When the comet's total energy is slightly less than zero, it must return to the sun. In this situation the comet's orbit is an ellipse, even though it may be a very extended ellipse. A comparison of an extended ellipse and a parabola is shown in Figure (14c). From this figure we can see that near the sun there is not much difference in the motion of a comet with zero or slightly negative total energy. The difference can be seen at a great distance, where the zero-energy comet continues to move away from the sun, but the slightly negativeenergy comet returns.

The circular, or nearly circular, motion of the planets is a limiting case of elliptical motion. The small circular orbits (Figure 14d) are occupied by planets that have large negative total energies. Thus the planets are tightly bound to the sun.

# Example 4 Escape Velocity

At what speed must a shell be fired from a super cannon in order that it escapes from the earth? Does it make any difference at what angle the shell is fired, so long as it clears all obstructions? (Neglect air resistance.)

**Solution**: If the shell is fired at a sufficiently great initial speed so that its total energy is positive, it will eventually escape from the earth, regardless of the angle at which it is fired (so long as it clears obstructions). To calculate the minimum muzzle velocity at which the shell can escape, we will assume that the shell has zero total energy, so that it barely escapes. When  $E_{total} = 0$  we have just after the shell is fired

$$0 = \frac{1}{2} \operatorname{mv}^2 - \frac{\operatorname{GM}_{e} \operatorname{m}}{\operatorname{r}_{e}}$$

which gives

$$v^2 = \frac{2GM_e}{r_e}$$
(55)

Putting in numbers

$$G = 6.67 \times 10^{-8} \text{ cm}^3/\text{gm sec}^2$$
  
 $M_e = 5.98 \times 10^{27} \text{ gm}$   
 $r_e = 6.38 \times 10^8 \text{ cm}$ 

we get

$$v^{2} = \frac{2 \times (6.67 \times 10^{-8} \text{ cm}^{3}/\text{gm sec}^{2})(5.98 \times 10^{27} \text{ gm})}{6.38 \times 10^{8} \text{ cm}}$$
$$= \frac{2 \times 6.67 \times 5.98}{6.38} \frac{10^{-8} \times 10^{27}}{10^{8}} \text{ cm}^{3} \text{gm/gm cm sec}^{2}$$
$$= 1.25 \times 10^{12} \text{ cm}^{2}/\text{sec}^{2}$$

 $v_{escape} = 1.12 \times 10^6 \text{ cm/sec}$ 

Converting this to more recognizable units, such as mi/ sec, we have

$$v_{escape} = 1.12 \times 10^6 \text{ cm/sec} \times \frac{1}{1.6 \times 10^5 \text{ cm/mi}}$$
  
= 7 mi/sec (11.2 km / sec)

This is also equal to 25,200 mi/hr, which is far faster than the initial velocity required to put a satellite in an orbit 100 mi high.

#### Exercise 12

Calculate the escape velocity required to project a shell permanently away from the moon ( $m_{moon} = 7.35 \times 10^{25}$  gm,  $r_{moon} = 1.74 \times 10^8$  cm).

#### Exercise 13

Once a shell has escaped from the earth, what must its speed be to allow it to escape from our solar system?

#### **Exercise 14**

Find the escape velocities from the earth and the moon, using the planetary units given on page 8-14.

## **BLACK HOLES**

A special feature of satellite motion we have just seen is that we can tell whether or not a satellite can escape simply by comparing kinetic energy with the gravitational potential energy. If the satellite's positive kinetic energy is greater in magnitude than the negative gravitational potential energy, then the satellite escapes, never to return on its own. This is true no matter how or from where the satellite is launched (provided it does not crash into something.)

So far we have limited our discussion to slowly moving objects where the approximate formula  $1/2 \text{ mv}^2$  is adequate to describe kinetic energy. We got the formula  $1/2 \text{ mv}^2$  back in Equation 7 by expanding  $\text{E} = \text{mc}^2$  to get

$$E = mc^{2} = \frac{m_{0}c^{2}}{\sqrt{1 - v^{2}/c^{2}}}$$

$$\approx m_{0}c^{2} + \frac{1}{2}m_{0}v^{2} \qquad (7)$$

$$(rest \ energy) \quad (kinetic \ energy)$$

The basic idea is that Einstein's formula  $E = mc^2$  gives us a precise formula for the sum of the rest energy and the kinetic energy. In the special case the particle is moving slowly, we can use the approximate formula for  $\sqrt{1 - v^2/c^2}$  to get the result shown in Equation 7.

For familiar objects like bullets, cars, airplanes, and rockets, the kinetic energy is  $1/2 \text{ m}_0 \text{v}^2$ , much, much, much smaller than the rest energy  $\text{m}_0\text{c}^2$ . The kinetic energy of a rifle bullet, for example, is enough to allow the bullet to penetrate a few centimeters into a block of wood. The rest energy of the bullet, if converted into explosive energy, could destroy a forest. In fact, one way to tell whether or not the approximate formula  $1/2 \text{ m}_0 \text{v}^2$  is reasonably accurate, is to check whether the kinetic energy is much less than the rest energy. If it is, you can use the approximate formula; if not, you can't.

We now finish our discussion of satellite motion by going to the opposite extreme, and consider the behavior of particles whose kinetic energy is much greater than their rest energy. Such a particle must be moving at a speed very close to the speed of light. We considered such a particle in Exercise 7 of Chapter 6. There we saw that electrons emerging from the Stanford linear accelerator travelled at a speed v = .9999999999875 c, and had a mass 200,000 times greater than the rest mass. For such a particle, almost all the energy is kinetic energy. In the formula  $E = mc^2$ , only one part in 200,000 represents rest energy.

Actually we wish to go one step farther, and discuss particles with no rest energy, particles that move *at* the speed of light. The obvious example, of course, is the photon, the particle of light itself.

From one point of view there is not much difference between an electron travelling at a speed .99999999999875 c with only 1 part in 200,000 of its energy in the form of rest energy, and a photon traveling at a speed c and no rest energy. Taking this point of view, we will take as the formula for the energy of a photon  $E = mc^2$ , and assume that this is pure kinetic energy.

Applying the formula  $E = mc^2$  to a photon implies that a photon has a mass  $m_p = E_p/c^2$ . We will now make the assumption that this mass  $m_{photon}$  is gravitational mass, and that photons have a gravitational potential energy  $-GMm_p/r$  like other objects. Our assumption, which is slightly in error, is that Newtonian gravity, which is a non relativistic theory, applies to particles moving near to or at the speed of light. It turns out that Einstein's relativistic theory of gravity gives almost the same answers, that we are seldom off by more than a factor of 2 in our predictions. Suppose we have a photon a distance r from a star of mass  $M_s$ . If the photon has a mass  $m_p$ , then the formula for the total energy of the photon, its kinetic energy  $m_pc^2$  plus its gravitational potential energy  $-GM_sm_p/r$  is

$$E_{tot} = m_p c^2 - \frac{GM_s m_p}{r}$$

Since  $m_p$  appears in both terms, we can factor it out (and also take out a factor of  $c^2$ ) to get

$$E_{tot} = m_p c^2 \left[ 1 - \frac{GM_s}{rc^2} \right]$$
(56)

Equation 56 applies only when the photon is outside the star, i.e., when the distance r is greater than the radius R of the star.

In most cases, the gravitational potential energy is much less than the kinetic energy of a photon, and gravity has little effect on the motion of the photon. For example, if a photon were grazing the surface of the sun (if r in Equation 56) were equal to the sun's radius  $R_{sun}$ ) we would have

$$E_{tot} = m_p c^2 \left[ 1 - \frac{GM_s}{R_s c^2} \right]$$
(57)

Putting in numbers  $M_s = 1.99 \times 10^{33} \text{gm}$ ,  $R_s = 6.96 \times 10^{10} \text{cm}$  we have

$$\frac{GM_s}{R_sc^2} = \frac{6.67 \times 10^{-8} \frac{cm^3}{gm sec^2} \times 1.99 \times 10^{33}gm}{6.96 \times 10^{10} cm \times (3 \times 10^{10})^2 \frac{cm^2}{sec^2}}$$
$$= .00000212$$

Thus

$$E_{tot} = m_p c^2 [1 - .00000212]$$
(58)

From Equation 58 we see that when a photon is as close as it can get to the surface of the sun, the gravitational potential energy contributes very little to the total energy of the photon, only 2 parts in a million.

However, suppose that the a star had *the same mass as the sun but a much, much smaller radius*. If it's radius  $R_s$  were small enough, the factor  $[1 - GM_s/R_sc^2]$  in Equation 58 would become negative, and a photon grazing the surface of this star would have a negative total energy. The photon could not escape from the star. No photons emerging from the surface of such a star could escape, and the star would cease to emit light.

Let us see how small the sun would have to be in order that it could no longer radiate light. That would happen when the factor  $[1 - GM_s/R_sc^2]$  is zero, when photons emerging from the surface of the sun have zero total energy. Putting in numbers we get

$$\frac{\mathrm{GM}_{\mathrm{s}}}{\mathrm{R}_{\mathrm{s}}\mathrm{c}^2} = 1 \tag{59}$$

$$R_{s} = \frac{GM_{s}}{c^{2}}$$
$$= \frac{6.67 \times 10^{-8} \frac{\text{cm}^{3}}{\text{cm sec}^{2}} \times 1.99 \times 10^{33} \text{cm}}{(3 \times 10^{10})^{2} \text{cm}^{2}/\text{sec}^{2}}$$

$$R_s = 1.48 \times 10^5 cm$$
  
= 1.48 kilometer (60)

Equation 60 tells us that an object with as much mass as the sun, confined to a sphere of radius less than 1.48 kilometers, cannot radiate light. Although we used the non relativistic Newtonian gravity in this calculation, Einstein's relativistic theory of gravity makes the same prediction. In discussions of black holes, one often sees a reference to the radius of the black hole. What is usually meant is the radius given by Equation 59, the radius at which light can no longer escape if a mass  $M_s$  is contained within a sphere of radius  $R_s$ .

Do black holes exist? Can so much mass be concentrated in such a small sphere? The question has been difficult to answer because black holes are hard to observe since they do not emit light. They have to be detected indirectly, from the gravitational pull they exert on neighboring matter. In the sky there are many binary star systems, systems in which two stars orbit about each other. In some examples we have observed a bright star orbiting about an invisible companion. Careful analysis of the orbit of the bright star suggests that the invisible companion may be a black hole. There is recent evidence that gigantic black holes, with the mass of millions of suns, exists at the center of many galaxies, including our own.

That a black hole cannot radiate light is only one of the peculiar properties of these objects. When so much matter is concentrated in such a small volume of space, the gravitational force becomes so great that other forces cannot resist the crushing force of gravity, and as far as we know, the matter inside the black hole collapses to a point—a zero sized or very, very small sized object. At the present time we do not have a good theory for what happens to the matter inside a black hole. (We will have more to say about black holes in later chapters.)

## Exercise 15

Studies of the motion of the stars in our galaxy suggests that at the center of our galaxy is a large amount of mass concentrated in a very small volume. For this problem, assume that a mass of 100 million suns is concentrated in the small volume. If this massive object is in fact a black hole, what is the radius from which light can no longer escape?

# A Practical System of Units

In the CGS system of units, where we measure distance in centimeters, mass in grams and time in seconds, the unit of force is the dyne (1 dyne = 1 gm cm/sec<sup>2</sup>) and the unit of energy is the erg (1 erg = 1 gm cm<sup>2</sup>/sec<sup>2</sup>). We have found the CGS system quite convenient for analyzing strobe photographs with 1 cm grids. But when we begin to talk about forces and particularly energy, the CGS system is often rather inconvenient. A force of one dyne is more on the scale of the force exerted by a fly doing push-ups than the kind of forces we deal with in the lab. A baseball pitched by Roger Clemens has a kinetic energy of over a million ergs and a 100 watt light bulb uses ten million ergs of electrical energy per second. The dyne and particularly the erg are much too small a unit for most every day situations.

In the MKS system of units, where we measure distance in meters, mass in kilograms and time in seconds, the unit of force is the newton and energy the joule. The force required to lift a 1 kilogram mass is 9.8 newtons (mg), and the energy of a Roger Clemens' pitch is over 10 joules. When working with practical electrical phenomena, the use of the MKS system is the only sensible thing to do. The unit of power, the watt, is one joule of energy per second. Thus a 100 watt light bulb consumes 100 joules of electrical energy per second. Volts and amperes are both MKS units, the corresponding CGS units are statvolts and esu, which are almost never used.

Where CGS units are far superior is in working with the basic theory of atoms, as for the case of the Bohr theory discussed in Chapter 36. This is because the electric force law has a much simpler form in CGS units. What we will do in the text from this point on is to use MKS units almost exclusively until we get through the chapters in electrical theory and applications. Then we will go back to the CGS system in most of our discussions of atomic phenomena.

# Chapter 11 Systems of Particles



So far in our applications of Newton's second law, we have treated objects as individual point particles. In some problems, this appeared to be a reasonable approximation. But in others, like the motion of an apple falling from a tree, it appeared to be a remarkable result that the entire earth could be treated as a point particle located at the center of the earth. Newton supposedly invented calculus to show that one could do this.

In this chapter we will look at ways to handle systems consisting of many particles. In particular, we will see that often the concept of **center of mass** allows us to treat a group of interacting particles as a single particle. This can lead to an enormous simplification of the analysis and a clearer understanding of the result.

#### **Diver - Movie**

For a student project, Tobias Hays was videotaped doing a series of dives, a few of which are shown in this movie. Working frame by frame, you can see that the diver's center of mass follows a parabolic trajectory. (It was actually more instructive to use a parabolic trajectory to locate the diver's center of mass in various positions.)

# **CENTER OF MASS**

To introduce the concept of center of mass, let us begin with some examples of mechanical systems that at first appear fairly complex, but which are greatly simplified when we focus our attention on the motion of the center of mass.

Consider the motion of the earth about the sun. To analyze the problem, we could first treat the sun as a point mass fixed at the origin and apply Newton's second law to the earth to determine the earth's elliptical orbit. On closer inspection, however, we note that the earth is not a point particle but an earth-moon system. A more accurate treatment of the problem requires us to consider two interacting particles both orbiting the sun.

With a computer program it is not too difficult to set up the earth-moon-sun system and directly solve for the motion of the earth and moon. Because the sun is so massive, it is still a good approximation to place the sun at rest at the origin of the coordinate system. We then have the earth subject to the gravitational force of the sun and the moon, and the moon experiencing the gravitational force of the sun and the earth. If we include all of these forces in the program, and start off with reasonable initial conditions, we will get the expected result that the earth goes about the sun in an elliptical orbit with a slight wobble, and the moon goes around the earth in a nearly circular orbit. If we plot the moon's orbit, exaggerating the orbit radius as in Figure (1), the orbit looks somewhat peculiar because it repre-



#### Figure 1

Motion of the moon around the sun. In this sketch, we have greatly exaggerated the size of the moon's orbit about the earth in order to show the epicycle like motion of the moon. sents circular motion about a moving center. It looks like the drawing of epicycles from a text on ancient Greek astronomy.

For a similar but much more difficult problem, consider the globular cluster shown in Figure (2). Globular clusters are fairly common objects in our galaxy. A typical globular cluster is a swarm of several million stars all attracting each other to form a single gravitationally bound body. We can think of it as a confined gas of stars, confined not by a bottle or a rubber balloon but by the gravitational attraction between the stars.

The globular clusters in our galaxy lie outside the main body of stars of the galaxy, typically orbiting around or through the galaxy. If we wished to calculate the orbit of a globular cluster, our first approximation would be to treat it as a point particle. To do better than that with a computer program, we might try to analyze the motion of each star as we did in the earth-moon-sun problem above. But the futility of doing this would soon become obvious. Each of the millions of stars interact not only with the gravitational force of the galaxy, but with each other. Each star is subject to millions of forces, and any direct computer calculation becomes impossibly complex. One might try to simplify the problem by considering a cluster of only a few



**Figure 2** *Globular cluster (NCG 5272).* 

hundred stars, but even then a lot of time on a super computer is needed for a meaningful prediction.

Despite all the forces involved, the motion of the cluster can easily be analyzed if we focus on the motion of the *center of mass* of the cluster. When we calculate the motion of the center of mass, all internal forces cancel, and we have to consider only the net force of the galaxy on the total mass of the stars. In the earth-moon-sun problem, the center of mass of the earth-moon system travels in an elliptical orbit around the sun. The earth and moon each orbit about this center of mass point. The calculation of the motion of the center of mass of the earth-moon is the same as calculating the motion of a point planet about the sun.

The idea of the center of mass is a familiar concept for it is a balance point. The center of mass of a long thin rod is the point where the rod balances on your finger. If you mark the balance point and throw the rod in the air, giving it a spin, the balance point follows a smooth parabolic trajectory while the rest of the rod rotates about the balance point as shown in Figure (3). The balance point or center of mass moves as if all the mass of the rod were concentrated at that point.

Although the idea of a center of mass or balance point is fairly straightforward, the formula for the center of mass of a collection of particles looks a bit peculiar at first. But when you get used to the formula, you will find that it is fairly easy to remember and leads to impressive simplifications when used in Newton's laws.



#### Figure 3

A meter stick tossed in the air rotates about its center of mass (balance point). The center of mass itself travels along the parabolic path of a point projectile.

#### **Center of Mass Formula**

Suppose that we have a collection of n particles  $m_1, m_2, \dots, m_n$ , as shown in Figure (4). The coordinate vector for particle  $m_1$  is  $\vec{R}_1$ , that for  $m_2$  is  $\vec{R}_2$ , etc. We define the *total mass* **M** of the collection of particles as simply the sum of the individual masses

$$M = m_1 + m_2 + \dots m_i = \sum_{i=1}^{n} m_i$$
 (1)

The *coordinate vector*  $\vec{R}_{com}$  of the center of mass point is then defined by the formula

$$M\vec{R}_{com} = \sum_{i} m_{i}\vec{R}_{i} \qquad formula for \\ center of mass \\ coordinate \qquad (2)$$

Since this is a vector equation, it can be written as the collection of the three scalar equations

$$MX_{com} = \sum_{i} m_{i} x_{i}$$
 (2a)

$$MY_{com} = \sum_{i} m_{i} y_{i}$$
(2b)

$$MZ_{com} = \sum_{i} m_{i} z_{i}$$
 (2c)



#### Figure 4

To calculate the center of mass of a collection of particles, you start by constructing the coordinate vector for each particle. You then use Equation 2 to calculate the center of mass coordinate vector. where  $X_{com}$ ,  $Y_{com}$  and  $Z_{com}$  are the x, y and z coordinates of the center of mass, and  $x_i$ ,  $y_i$ , and  $z_i$  are the x, y, z coordinates of the i-th particle.

To see that Equation 2 does give the balance point of a collection of particles, let us consider the simple case of a horizontal massless rod of length l with two masses  $m_1$  and  $m_2$  at each end, as shown in Figure (5). If  $m_1$  is placed at the origin of the coordinate system, then Equation 2a gives

$$M X_{com} = m_1 \times 0 + m_2 l$$
$$X_{com} = \frac{m_2 l}{M} = \frac{m_2 l}{m_1 + m_2}$$

If  $m_1$  and  $m_2$  are the masses of two children on a seesaw of length *l*, then the pivot should be placed a distance  $X_{com}$  from  $m_1$  for the children to balance. (As a quick check, if  $m_1 = m_2$ , then  $X_{com} = l/2$  which is obviously correct.)

[If you want to calculate the center of mass of a continuous object like an irregularly shaped sheet of plywood, you can mark the plywood off into small sections of mass  $dm_i$  located at  $\vec{r}_i$ . Then M is the mass of the whole sheet of plywood, and the x coordinate of the center of mass is

$$MX_{com} = \sum_{i} x_{i} dm_{i} \rightarrow \int x dm$$

where  $x_i$  is the x coordinate of  $\vec{r}_i$ . You can then replace the sum over the  $dm_i$  by an integral. This is a typical problem treated in an introductory calculus course and will not be discussed further here.]



**Figure 5** *Calculating the center of mass for two particles.* 

#### Exercise 1

The formula for center of mass looks like it depends on where you place the origin of the coordinate system. To see that this is not true recalculate  $X_{com}$  for the two masses in Figure (5), placing the origin at  $m_2$ , and show that the pivot point comes out at the same place on the rod.

#### Exercise 2

Show that the center or mass of the earth-moon system is located inside the earth.

#### **Exercise 3**

An ammonia molecule consists of a nitrogen atom and three hydrogen atoms located on the corners of a tetrahedron, as shown in Figure (6). Locate the center of mass of the ammonia molecule.

## **Dynamics of the Center of Mass**

Let us now see how the concept of center of mass can be used to handle the dynamic behavior of a system of particles. Suppose we have a system with n particles. The formula for their center of mass coordinate  $\vec{R}_{com}$ is from Equation 2

$$M\vec{R}_{com} = m_1\vec{R}_1 + m_2\vec{R}_2 + \dots + m_n\vec{R}_n$$
 (2)

If we differentiate Equation 2 with respect to time, and note that the velocity  $\vec{V}_{com}$  is given by  $d\vec{R}_{com}/dt$ , we get

$$\vec{MV}_{com} = m_1 \vec{v}_1 + m_2 \vec{v}_2 + \dots + m_n \vec{v}_n$$
 (3)

where  $\vec{v}_i = d\vec{R}_i/dt$  is the velocity of the i-th particle.



**Figure 6** *Structure of the ammonia molecule.* 

Equation 3 already has an interesting interpretation. Since  $m_1 \vec{v}_1$  is the linear momentum of particle  $1, m_2 \vec{v}_2$  that of particle 2, etc., we see that  $M \vec{V}_{com}$  is equal to the vector sum of the linear momenta of all the particles under consideration. We will come back to this point when we discuss the concept of linear momentum in more detail.

Differentiating Equation 3 with respect to time, noting that  $\vec{A}_{com} = d\vec{V}_{com}/dt$  is the acceleration of the center of mass point, we get

$$\vec{MA}_{com} = m_1 \vec{a}_1 + m_2 \vec{a}_2 + \dots + m_n \vec{a}_n$$
 (4)

where  $\vec{a}_i = d\vec{v}_i/dt$  is the acceleration of the i-th particle.

The final step is to use Newton's second law to replace  $m\vec{a}_i$  by the vector sum of the forces acting on particle i. Calling this sum  $\sum \vec{F}_i$ , we get

$$\vec{MA}_{com} = \sum \vec{F}_1 + \sum \vec{F}_2 + \dots + \sum \vec{F}_n \qquad (5)$$

Equation 5 tells us that  $\vec{MA}_{com}$  is equal to the sum of every force acting on all the particles.

If we wish to apply Equation 5 to the motion of something as complex as a globular cluster, it looks like we are still in trouble because, as we have mentioned, each of the millions of stars in the cluster interacts with all the other stars in the cluster. Acting on each star is the gravitational force of the galaxy plus the millions of forces exerted by the other stars.

But Newton's law of gravity provides an enormous simplification. When two objects interact gravitationally, they exert equal and opposite forces on each other as shown in Figure (7). In that case  $\vec{F}_{g1} = -\vec{F}_{g2}$  or if we add these two forces of interaction we get



**Figure 7** *Two objects exert equal and opposite forces on each other.* 

Let us now write out Equation 5 as it might be applied to the globular cluster:

$$M\vec{A}_{com} = \left(\vec{F}_{G1} + \vec{F}_{21} + \vec{F}_{31} + \cdots\right) + \left(\vec{F}_{G2} + \vec{F}_{12} + \vec{F}_{32} + \cdots\right) + \cdots$$
(6)

where  $\vec{F}_{G1}$  is the force of the galaxy on star #1,  $\vec{F}_{21}$  is the force of star #2 on star #1,  $\vec{F}_{31}$  the force of star #3 on star #1, etc. In the next collection of terms we have  $\vec{F}_{G2}$ , the force of the galaxy on star 2,  $\vec{F}_{12}$  the force of star #1 on star #2, etc.

Rearranging the order in which we write the forces, we get

$$\begin{split} \mathbf{M}\vec{A}_{com} &= \left(\vec{F}_{G1} + \vec{F}_{G2} + \cdots \vec{F}_{Gn}\right) \\ &+ \left(\vec{F}_{21} + \vec{F}_{12}\right) \\ &+ \left(\vec{F}_{31} + \vec{F}_{13}\right) + \cdots \end{split} \tag{7}$$

In Equation 7, we have separated the external forces  $\vec{F}_{G1} + \vec{F}_{G2} + \cdots$  exerted by the galaxy on the individual stars from the internal forces like  $\vec{F}_{12}$  and  $\vec{F}_{21}$  between stars in the cluster. All the internal forces can be grouped in pairs, like  $(\vec{F}_{12} + \vec{F}_{21})$ , the force of star #2 on star #1 plus the force of star #1 on star #2. Because these are equal and opposite forces, all the pairs of internal forces cancel (over a trillion pairs for an average cluster), and we are simply left with the vector sum of the external forces. In our cluster example, the vector sum of all the external forces is just the net force  $\vec{F}_{ext}$  the galaxy exerts on the cluster, and we are left with the fantastically simple result

$$\mathbf{M}\vec{A}_{com} = \vec{F}_{ext} \qquad \begin{array}{c} equation for \\ center of \\ mass motion \end{array}$$
(8)

Equation 8 tells us that the center of mass of the globular cluster moves exactly as if the cluster were a single mass point of mass M equal to the total mass of the cluster subject to a single force  $\vec{F}_{ext}$  equal to the total gravitational force exerted by the galaxy on all the stars in the cluster. This remarkable result explains

why we can often represent a complex system by a single mass point in the analysis of the system's behavior.

When this result is applied to the earth-moon system shown in Figure (8), we have the following picture. When we calculate the motion of the center of mass of the earth and moon about the sun, the internal forces between the earth and moon cancel, and we are left with

$$\left(\mathbf{M}_{\text{earth}} + \mathbf{M}_{\text{moon}}\right) \vec{\mathbf{A}}_{\text{com}} = \left(\vec{\mathbf{F}}_{\text{sun on earth}} + \vec{\mathbf{F}}_{\text{sun on moon}}\right)$$

where  $\dot{F}_{sun on earth}$  is the force of the sun on the earth, and  $\vec{F}_{sun on moon}$  that of the sun on the moon. The center of mass moves with the same acceleration as a point particle of mass ( $M_{earth} + M_{moon}$ ) subject to the total force the sun exerts on the two. This results in an elliptical orbit for the center of mass.

## **Exercise** 4

Two air carts of equal mass are connected by a spring as shown in Figure (9). A small black marker is placed at the center of the spring which remains at the center of mass of the carts. One of the carts is given a shove to the right so that the two carts move off to the right in an undulating drift. Describe the motion of the black marker.



Figure 8

Forces on the earth and moon, as they go around the sun. (The force of the sun on the earth is much larger than the other three forces shown.)

# **NEWTON'S THIRD LAW**

In our analysis of the motion of a globular cluster, the great simplification came when all the internal forces canceled in pairs in Equation 7, and we were left with the result that the acceleration of the center of mass point was determined solely by the external forces acting on the swarm of stars. The cancellation occurred because the gravitational attraction between two objects is equal in magnitude and oppositely directed.



What about other forces? What if two stars are in collision? Is the force between them still equal and opposite? If not, then we would have to take internal forces into account when predicting the motion of the center of mass? Unable to believe that this would happen, Newton proposed that when two bodies interact, then the force between them is always equal in magnitude and oppositely directed, no matter what forces are involved. This assumption is known as **Newton's Third Law of Mechanics**. The third law guarantees that internal forces cancel in pairs and that center of mass motion is determined only by external forces.

[Newton's first law, as we have mentioned, is that in the absence of any external forces, an object will move with uniform motion. Although there is a direct consequence of the second law  $\vec{F} = m\vec{a}$ , Newton explicitly stated the result, because it was not such an obvious idea in Newton's time, when a horse and buggy was the smoothest ride available.

In a traditional course in Newtonian mechanics, Newton's three laws are presented at the beginning as basic postulates and everything else is derived from them. This was a logical approach for over 200 years during which time there were no known exceptions to these laws. But with the discovery of special relativity in 1905 and quantum mechanics in 1923, we now know that Newton's laws are an approximate set of equations which apply with great accuracy to objects like stars, planets, cars and baseballs, but which have to be significantly modified when we consider objects moving at speeds near the speed of light, and which fail completely on a subatomic scale.]

#### CONSERVATION OF LINEAR MOMENTUM

In Chapter 6, after introducing the recoil definition of mass, we saw that the quantity  $m_1 \vec{v}_1 + m_2 \vec{v}_2$  remained unchanged when the two objects were recoiled from each other.



We proposed that this was one example of a more general conservation law-the conservation of linear momentum. Now, using Newton's third law, we can explicitly demonstrate how the law of conservation of momentum applies to objects obeying Newton's laws.

Our discussion begins with Equation 3, reproduced below as Equation 9, that was obtained by differentiating the formula for the center of mass of a system of particles. The result was

$$\overrightarrow{WV}_{com} = m_1 \overrightarrow{v}_1 + m_2 \overrightarrow{v}_2 + \dots + m_n \overrightarrow{v}_n$$
  
=  $\overrightarrow{P}_{total}$  (9)

where  $\vec{P}_{total} = m_1 \vec{v}_1 + \dots + m_n \vec{v}_n$  is the vector sum of the momenta of all the particles in the system. We will call this the total momentum of the system. We see from Equation 9 that this total momentum is simply the total mass M times the velocity of the center of mass.

Differentiating Equation 9 with respect to time, as we did in going from Equation 3 to 4 we get

$$\frac{\vec{dP}_{total}}{dt} = m_1 \vec{a}_1 + m_2 \vec{a}_2 + \dots + m_n \vec{a}_n \qquad (10)$$

This is essentially Equation 4, except that we are replacing  $\vec{MA_{com}}$  by  $dP_{total}/dt$ .

Using the fact that  $m_1 \vec{a}_1$  is the vector sum of all the forces acting on particle 1,  $m_2 \vec{a}_2$  the sum acting on particle 2, etc., we get

1

$$\frac{d\vec{P}_{total}}{dt} = \begin{cases} sum of all the forces \\ acting on all the particles \end{cases}$$
(11)

Now use Newton's third law to cancel all the internal forces, and we are left with

$$\frac{d\vec{P}_{total}}{dt} = \vec{F}_{ext} \qquad Newton's second law for a system of particles} (12)$$

where  $\vec{F}_{ext}$  is the vector sum of all the external forces acting on the system.

If we have an isolated system with no external forces acting on it (if our globular cluster were drifting through empty space), then

$$\frac{d\vec{P}_{total}}{dt} = 0 \qquad \begin{array}{c} for \ an \\ isolated \\ system \end{array}$$
(13)

Equation 13 is our desired statement of the law of conservation of linear momentum. In words it says that the total linear momentum of an isolated system is conserved-does not change with time. (Linear momentum is also conserved if there are external forces but their vector sum is zero. For example, a cart on an air track experiences the downward force of gravity and the upward force of the air, but these forces exactly cancel.)

In deriving Equation 13, we had to use Newton's hypothesis that all internal forces cancel in pairs. And we also needed Newton's second law to relate  $d(m\vec{v_i})/dt$ to the vector sum of the forces acting on the i-th particle. However, the law of conservation of linear momentum is known to apply on a subatomic scale where the concepts of force and acceleration loose their meaning. This suggests that our derivation of momentum conservation is somehow backwards. A more logical route is to assume conservation of linear momentum as a basic principle and derive the consequences.

[To see how such a derivation would look, consider an isolated system where there are no external forces. We get from Equation 11

$$\frac{d\vec{P}_{total}}{dt} = \begin{array}{c} \text{vector sum of all the} \\ \text{internal forces acting} \\ \text{on the particles} \end{array}$$
(14)

But conservation of linear momentum requires that the linear momentum of an isolated system be conserved. Thus  $d(P_{Total})/dt$  must be zero, and therefore the vector sum of all the internal forces must be zero. This must be true no matter what kind of forces are involved. Newton's third law is a bit more restrictive in that it requires the internal forces to cancel in pairs. The cancellation in pairs is the simplest picture, but not necessarily required for conservation of linear momentum.]

## Momentum Version of Newton's Second Law

In our discussion of Newton's second law, we have consistently assumed that the mass of a particle was constant, so that we could take the m outside the derivative. For example, in going from Equation 3 to Equations 4 and 5 in the center of mass derivation, we used the following steps to relate the rate of change of momentum of the i-th particle  $d(m_i \vec{v}_i)/dt$  to the total force  $\vec{F}_i$  acting on the particle

$$\frac{d(m_i \vec{v}_i)}{dt} = \frac{m_i d \vec{v}_i}{dt} = m_i \vec{a}_i = \vec{F}_i$$
(15)

If we take  $\vec{F} = m\vec{a}$  to be the basic form of Newton's second law, we have to assume m is constant in order to get  $m\vec{a}$  and then  $\vec{F}$ .

Newton recognized that a slight generalization of the second law would make it unnecessary to assume that a particle's mass was constant. He actually expressed the second law in the form

$$\vec{F} = \frac{d\vec{p}}{dt} = \frac{d}{dt} (m\vec{v})$$
(16)

where  $\vec{F}$  is the total force acting on a particle of mass m, moving with a velocity  $\vec{v}$ . In the special case that m is constant, then Newton's second law becomes

$$\vec{F} = \frac{d(m\vec{v})}{dt} = m\frac{d\vec{v}}{dt} = m\vec{a}$$
 (m = constant) (17)

Thus we should view the Equation  $\vec{F} = m\vec{a}$  as a special case of the more general law  $\vec{F} = d\vec{p}/dt$ .

The momentum form of Newton's second law is advantageous if we are considering problems like that in the following exercise where momentum is being transferred to an object at a known rate and we wish to determine the effective force. A more basic application is for relativistic problems where mass changes with velocity. Then we must use the momentum form of the law in order to account for the mass change. It is interesting that Newton had the insight to present the second law in a form that would handle Einstein's theory 200 years later.

#### Exercise 5

A boy is washing the door of his father's car by squirting a hose at the door. Assume that the water comes out of the hose at a rate of 20 kilograms (liters) per minute at a speed of 12 meters/second. When the water hits the door it dribbles down the side. What force, in newtons, does the water exert on the door?

(To solve this, simply calculate the amount of momentum per second that the water brings to the door.)

## COLLISIONS

In our every day experience, collisions are something we usually try to avoid, whether it is running into a door or an automobile accident. Hitting a baseball is an obvious exception. In physics, collisions turn out to play an extremely important role, particularly in the study of elementary particles. For example, the atomic nucleus was discovered as a result of experiments involving the collision between  $\alpha$  particles and atoms in a gold foil.

Collisions generally happen rapidly, and one is not used to observing what happens during a collision. It is usually a before and after scene, what was the situation before the collision, and what did things look like after. In most physics experiments like those involving elementary particles, that is all we can observe. However we will begin our discussion of collisions with an experiment that is explicitly designed to allow us to study the situation during the collision.

In this experiment, an air cart moving down an air track collides with a force detector mounted at the end of the track. Rather than have the metal cart bounce off the metal arm of the force detector, we slow the collision down by mounting a stretched rubber band on the end of the cart. With the rubber band colliding with the force detector, it takes several milliseconds for the collision to occur and the cart to reverse directions. During this time we can record both the force the cart exerts on the force detector and the velocity of the cart. Using the momentum form of Newton's second law, we will find that there is a particularly simple way to analyze the collision in terms of the concept of *impulse*.

When collisions are either too rapid or on too small a scale to be observed directly, we can almost always apply the law of conservation of momentum to analyze the results. In elementary particle collisions, both energy and momentum may be conserved. In some introductory physics lab experiments, momentum is conserved during the collision and energy after. We will see how this lets us make detailed predictions about the behavior of the objects involved.

#### Impulse

An overview of the air cart force detector experiment is shown in Figure (10). Focusing on the momentum involved, we see that the cart is initially moving down the track with a momentum  $\vec{p}_i$  as shown in Figure (10a). In Figure (10b) it collides with the force detector, and in (10c) is moving back up the track with a momentum  $\vec{p}_f$ . The net effect of the collision is to change the cart's momentum from  $\vec{p}_i$  to  $\vec{p}_f$ .



#### Figure 10

Collision of an aircart with the force detector. During the collision, the force detector first has to push on the cart to stop it, and then give an essentially equal push to move it out. During the collision itself, the force detector is exerting a force  $\vec{F}(t)$  on the cart. The force  $\vec{F}(t)$  acts for only a short time, but can be measured in detail by the force detector. To relate this force to the observed change in the momentum of the cart, we start with Newton's law in the form

$$\vec{F}(t) = \frac{d\vec{p}}{dt}$$
 (16 repeated)

Multiplying through by dt gives

$$\vec{F}(t)dt = d\vec{p} \tag{18}$$

Now integrate both sides of this equation from a time  $t_i$  before the collision, when the cart momentum was  $\vec{p}_i$ , to a time  $t_f$  after the collision, when the cart momentum was  $\vec{p}_f$ . We get

$$\int_{t_{i}}^{t_{f}} \vec{F}(t) dt = \int_{p_{i}}^{p_{f}} d\vec{p} = \vec{p}_{f} - \vec{p}_{i}$$
(19)

Since  $(\vec{p}_f - \vec{p}_i)$  is the change in the momentum of the cart as a result of the collision, we get

impulse of the  $\int_{t_i}^{t_f} \vec{F}(t) dt =$  change in the momentum of the air cart (20)

This integral of  $\vec{F}(t)$  over the time of the collision is called *the impulse* of the force  $\vec{F}(t)$ . The force exerted by the force detector alters the momentum of the cart, and how much it alters it is equal to the impulse of the force.



**Figure 11** *Calibrating the force detector.* 

In the air cart experiment, we will record the force  $\vec{F}(t)$  from the output of the force detector, and directly compare the integral of that force with the change in the momentum of the cart.

Note that both  $\vec{p}_f$  and  $-\vec{p}_i$  are directed to the right. Thus the magnitude of  $\vec{p}_f - \vec{p}_i$  is equal to the numerical sum  $p_f + p_i$ 

$$\left|\vec{p}_{f} - \vec{p}_{i}\right| = \left|\vec{p}_{f}\right| + \left|\vec{p}_{i}\right|$$

As a result, the impulse supplied by the force detector has a magnitude  $p_f + p_i$  or about  $2p_i$  if the cart comes out at the same speed it went in. The force detector supplies  $2p_i$  because one  $p_i$  is required to stop the cart, and the other is required to shove it out again.

## **Calibration of the Force Detector**

The force detector is designed to put out a voltage that is proportional to the force exerted on the detector beam. To convert this voltage reading to a force measurement, we have to calibrate the force detector. This is easily done by running a string from the air cart, over a pulley and down to some weights as shown in Figure (11). As we add weights to the string we increase the force the cart exerts on the beam.

Figure (12) shows the output of the force detector as we added a series of 20 gm weights. Adding 3 weights changes the voltage by 42.8 millivolts (mV), thus each weight changes the output voltage by 42.8/3 = 14.3 mV. (One millivolt =  $10^{-3}$  volts.) Each added weight corresponds to an increase of the force by  $\Delta F = mg = 20 \text{ gm} \times 980 \text{ cm/s}^2 = 19600 \text{ dynes}$ . Thus the factor for converting from millivolts output for this force detector, to dynes of force is



**Figure 12** *Voltage output when 20 gram weights are added.* 

$$\frac{\text{conversion}}{\text{factor}} = \frac{19600 \text{ newtons}}{14.3 \text{ millivolts}}$$

$$= 1370 \frac{\text{dynes}}{\text{millivolts}}$$
(21)

#### The Impulse Measurement

One way to set up the collision experiment is shown in Figure (13). To soften the collision we have added a metal bracket to the cart and stretched a rubber band across the open end. Adjusting the tension in the rubber band allows us to change the length of time during which the collision occurs.

Figure (14) shows a fairly typical output of the force detector. There is no force until the rubber band reaches the detector. The force then increases and then decreases symmetrically, and becomes 0 when the cart leaves. Using our calibration factor of 1370 dynes/mV, we have graphed the force, in dynes, as a function of time in Figure (15).

The impulse of this force is the integral of the force curve from the time  $t_1$  that the rubber band gets to the force detector, to  $t_2$  when it leaves. Since we do not



Figure 13

A rubber band is used to soften the collision.



Figure 14

*Output of the force detector. There is zero force before the cart arrives and after it leaves.* 

have a formula for the curve, we cannot do the integral analytically. Instead, we have to use some graphical technique to find the area under that curve. One way is to superimpose the curve on graph paper and count the squares underneath. A slightly less accurate way we will use is to construct a triangle whose area is, to our best estimate, equal to the area under the curve. We have done this with the dashed line triangle seen in Figure (15). We have adjusted the triangle so that the extra area at the top matches the area lost at the sides.

The area of a triangle is  $(1/2 \text{ base} \times \text{altitude})$ . In Figure (15), the base of the triangle is 3.00-2.60 = 0.40 seconds. The height is seen to be close to 56800 dynes. Thus the area is

$$\frac{\text{area of}}{\text{triangle}} = \frac{1}{2} \times (0.40 \text{ sec}) \times (56800 \text{ dynes})$$

$$= 11360 \text{ dyne seconds}$$
(22)

Equation 22 is our result for the impulse of the force F(t).

Although the force detector measures the magnitude of the force F(t) that the cart exerts on the detector, by Newton's third law, this should be equal in magnitude but oppositely directed to the force exerted by the detector on the cart. Thus the magnitude of the impulse calculated in Equation 22 should be equal to the magnitude of the change in the momentum of the cart, as a result of the collision.



#### Figure 15

Force versus time graph. The area under the force curve is about equal to the area of the triangle we drew.

#### **Exercise 6**

(a) What is the direction of the force exerted by the force detector on the cart?

- (b) What is the direction of the vector  $\Delta \vec{p} = \vec{p}_f \vec{p}_i$ ?
- (c) How do these two directions compare?

#### **Change in Momentum**

To measure the momentum of the cart before and after the collision, we mounted a 10 cm long sail on the top of the cart as shown in the side view of Figure (16). Mounted above the track is a light source and a photodetector seen in the top view. When the sail on the cart interrupts the light beam, there is an abrupt change in the voltage output by the photodetector. The lower, dashed curves in Figures (14) and (17) are from the output of the photodetector. In Figure (17a) we are measuring the length of time the sail took to pass by the photocell on the way down to the force detector. We see that this time was 400 milliseconds or .400 seconds. Thus the velocity of the cart on the way down was

$$v_i = \frac{10 \text{ cm}}{.400 \text{ sec}} = 25 \frac{\text{cm}}{\text{sec}}$$
 (23)



#### Figure 16

One way to measure the velocity of the aircart is to mount a 10 cm long sail is mounted on top of the aircart. While the sail is interrupting the light beam, there is a change in the output voltage of the photo detector. On the way back, we find from Figure (17c) that the sail took 412 milliseconds or .412 seconds to pass the photodetector. Thus the final speed of the cart was

$$v_{\rm f} = \frac{10 \, \rm cm}{.412 \, \rm sec} = 24.27 \frac{\rm cm}{\rm sec}$$
 (24)

The cart, sail and rubber band apparatus had a total mass of  $m_{cart}$  which was measured to be

$$m_{cart} = 227 \text{ gm} \tag{25}$$



#### Figure 17a

Simultaneous recording of both the voltage output from the force detector (solid line) and the photo detector (dashed line).

**Figure 17b,c** *Measuring the length of time the sail took to go past the photodetector. The 10 cm sail took 400 milliseconds to pass on the way in (upper curve) and 412 milliseconds on the way out. We see it slowed down a bit.* 

(Another way to determine the speed of the aircart is to tilt the airtrack and release the cart from a known height.)



2.6 2.8 3.0 3.2 3.4

Thus the initial and final momenta have magnitudes

$$p_{i} = mv_{i} = 227 \text{ gm} \times 25.0 \frac{\text{cm}}{\text{sec}}$$
$$= 5680 \frac{\text{gm cm}}{\text{sec}}$$
$$p_{f} = mv_{f} = 227 \text{ gm} \times 24.27 \frac{\text{cm}}{\text{sec}}$$
$$= 5510 \frac{\text{gm cm}}{\text{sec}}$$

The magnitude  $\Delta p$  of the change in momentum is the sum of these two values

$$\Delta p = p_i + p_f = 5680 \frac{\text{gm cm}}{\text{sec}} + 5510 \frac{\text{gm cm}}{\text{sec}}$$

$$change in the
linear momentum
= \Delta p = 11200 \frac{\text{gm cm}}{\text{sec}}$$
(26)

From Equation 22, we saw that the total impulse supplied by the force detector was

total impulse  
from the  
force detector = 
$$\int_{t_i}^{t_f} \vec{F}(t) dt = 11360$$
 dyne sec

(27)

We see that to within a quite reasonable experimental error, the total impulse supplied by the force detector equals the change in linear momentum of the aircart.

#### Exercise 7

Explain why  $\Delta p$  in Equation 26 is the sum of the magnitudes  $p_i$  and  $p_f$ .

#### Exercise 8

Show that the dimensions of impulse (dyne seconds) are the same as momentum (gm cm/sec).

#### **Exercise 9**

How much energy, in joules, did the cart lose in the collision with the force detector? What percentage of the cart's initial kinetic energy was this?

## Momentum Conservation during Collisions

In our force detector experiment we used a rubber band to slow down the collision so that we could do a more accurate analysis of the impulse. Even so the impulsive force  $\vec{F}(t)$  acted for only a very short time. The most important point of the experiment is that, no matter how short the time is, the impulse, the time integral of  $\vec{F}(t)$  is equal to the total change in momentum. If we had let the metal end of the cart strike the force detector, the collision would have taken much less time, but the force would have been much greater. The integral of the larger force over the shorter time would still equal the change in the momentum of the cart.

Suppose that, instead of an air cart colliding with a force detector, we had two air carts colliding with each other. During the collision they would by Newton's third law, exert equal and opposite forces  $\vec{F}(t)$  on each other. Thus they would exert upon each other equal and opposite impulses  $\int \vec{F}(t) dt$ . As a result, the momentum gained by one cart would precisely be equal to the momentum lost by the other. The net result is conservation of momentum during the collision.

Now consider a slightly more complex situation. Suppose I throw a red billiard ball up in the air, and you throw a blue billiard ball, and the two balls collide before landing.

During this collision, more forces are involved. There is the force of the red billiard ball on the blue one, the force of the blue billiard ball on the red one, and there are the gravitational forces acting on both. To study the change in the momenta of the balls, it appears that we must now account for all four forces at once.

However there is something special about the impulsive forces found in collisions. These forces are usually very large but **act for a very short time**. During this short time the collision forces are usually much larger than external forces like gravity. So much larger, in fact, that we can usually neglect external forces during a collision. Since the collision forces conserve linear momentum, we get the result that linear momentum is conserved during a collision even if external forces are present. The only exception would be if the collision is so slow that the external forces have time to act and change the system's momentum during the collision. This is usually not the case.
## **Collisions and Energy Loss**

While we can use conservation of linear momentum in the analysis of a collision, often energy conservation is not applicable. If the objects are deformed or give off heat, light or sound, energy escapes in ways that are difficult to measure.

In some situations, however, we can use momentum conservation to figure out how much energy must be "*lost*" to deformation, heat, sound, etc. Suppose, for example, we hang a steel ball of mass M from a string as shown in Figure (18), and throw a putty ball of mass m at the steel ball. The putty ball is initially moving at a speed  $v_i$ , then hits and sticks to the steel ball. The two move off together at speed  $v_f$ .

Even though energy is "lost" when the putty ball squashes up against the steel ball, momentum is conserved during the collision. The initial momentum is all carried by the putty ball, and is thus

$$p_i = mv_i \quad \frac{initial}{momentum}$$
 (28)

After the collision the two move off together at a speed  $v_f$  and thus have a total momentum

$$\mathbf{p}_{\mathbf{f}} = (\mathbf{m} + \mathbf{M})\mathbf{v}_{\mathbf{f}} \tag{29}$$

Since momentum is conserved, we have

$$p_{i} = p_{f}$$

$$mv_{i} = (m+M)v_{f}$$
(30)



Figure 18

Collision of a putty ball with a steel ball.

Solving for the final speed  $v_f$ , we get

$$\mathbf{v}_{\mathrm{f}} = \left(\frac{\mathrm{m}}{\mathrm{m}+\mathrm{M}}\right) \mathbf{v}_{\mathrm{i}} \tag{31}$$

We can now calculate the amount of energy that must be dissipated in this collision. The initial energy  $E_i$  is the kinetic energy of the putty ball

$$E_i = \frac{1}{2}mv_i^2 \tag{32}$$

The final energy  $E_{f}$  is the kinetic energy of the two together

$$E_{f} = \frac{1}{2}(m+M)v_{f}^{2}$$
 (33)

The energy "lost", which must have gone into deforming the putty ball, is

$$E_{lost} = E_{i} - E_{f}$$
  
=  $\frac{1}{2}mv_{i}^{2} - \frac{1}{2}(m+M)v_{i}^{2}$  (34)

Using Equation 31 for  $v_f$  in Equation 34 gives

$$E_{lost} = \frac{1}{2}mv_{i}^{2} - \frac{1}{2}(m+M)\left(\frac{mv_{i}}{m+M}\right)^{2}$$
  
$$= \frac{1}{2}mv_{i}^{2} - \frac{1}{2}\left(\frac{m^{2}}{m+M}\right)v_{i}^{2}$$
  
$$= \frac{1}{2}mv_{i}^{2}\left(1 - \frac{m}{m+M}\right)$$
  
$$E_{lost} = \frac{1}{2}mv_{i}^{2}\left[\frac{M}{m+M}\right]$$
(35)

It may be somewhat surprising that even though we may not have the slightest idea how energy is lost during the deformation of the putty, we can calculate precisely how much energy this uses.

#### Exercise 10

Check that Equation 35 gives reasonable results for the two special cases M = 0 and  $M = \infty$  (for  $m \neq 0$ ).

#### Exercise 11

A 500 gram steel ball is suspended from a string as shown in Figure (18). It is struck by a putty ball that sticks, and half the putty ball's initial kinetic energy is lost in the collision. What is the mass of the putty ball? After the putty ball has collided with the steel ball in our preceding example, the two will rise together to a maximum height h and final angle  $\theta_f$  before swinging back down again. This is illustrated in Figure (19). We can predict this height h by applying energy conservation after the collision. The kinetic energy  $E_f$  just **after the collision** is transformed into gravitational potential energy (m + M)gh at the top of the swing, giving

 $\frac{1}{2}(m+M)v_f^2 = (m+M)gh$ 

or

$$h = \frac{v_f^2}{2g}$$
(36)

If you want to calculate the final angle  $\theta_f$ , you use the fact that  $h = \ell - (\ell \cos \theta_f)$  as can be seen from Figure (19).

This example of the steel and putty ball is essentially equivalent to the ballistic pendulum discussed in Exercise 12. In that problem a block of wood of mass M is suspended from two strings as shown in Figure (20). A bullet of mass m is fired into the block and the block with the bullet stuck inside rises to a height h as shown.



just after collision

#### Figure 19

After the collision, energy is conserved. This allows us to calculate how high the ball rises.

You can assume that momentum is conserved during the collision, energy afterward, and from a measurement of the height h, determine the speed  $v_i$  of the bullet before it hit the block. This provides a rather simple, inexpensive way to measure the speed of bullets.

From the prospective of an introductory physics course, the ballistic pendulum experiment clearly distinguishes the use of momentum conservation and energy conservation. During the collision, momentum is conserved but energy is not. During the rise up to a height h, energy is conserved but momentum is not. To analyze such problem, you must develop an understanding of when you can apply the conservation laws and when you cannot. (For a safer ballistic pendulum demonstration, see Figure (21) on the next page.)

#### Exercise 12

(a) A bullet of mass  $m_b$  traveling at a speed  $v_b$  is fired into a block of wood of mass M hanging at rest as shown in Figure (20). The combined block and bullet rise to a height h. Find a formula for the speed  $v_b$  of the bullet.

(b) For part (a), suppose the bullet's mass is  $m_b = 10$  gm, the block of wood has a mass M = 1000 gm, and the final height h is 12 cm. What was  $v_b$  in cm/sec?



**Figure 20** Ballistic pendulum experiment.

#### **Exercise 13**

In a lecture demonstration that is safer to perform than the above ballistic pendulum experiment, a wastebasket is suspended from two cords and a pillow is placed inside the wastebasket, as shown in Figure (21).

Various members of the class are selected to throw a softball into the wastebasket. A scale is constructed to indicated how fast the ball was thrown. If the mass m of the softball is 200 gm, the mass M of the wastebasket and pillow is 1000 gm, and the length of the suspension cords are 2 meters from the ceiling to the center of the wastebasket, determine the distance (x) that the end of the basket travels if the ball is thrown at 40 miles/hour.

## **Collisions that Conserve Momentum and Energy**

When a bullet plows into a block of wood, a considerable amount of energy goes into deforming the wood and bullet. On the other hand if two hardened steel ball bearings collide, almost all the energy stays in the form of kinetic energy of the particles and very little is "lost" as heat, sound, and the deformation of the objects. When no energy is lost this way, we say that the collision is *elastic*. If energy is lost, then the collision is called *inelastic*.



**Figure 21** A safer ballistic pendulum experiment for classroom demonstrations.

On an atomic and subatomic scale we do not have the usual sound and friction, and small deformations may not be allowed. In these circumstances there is no way for energy to become "lost" and the resulting collisions are truly elastic. Thus in the study of the collisions of atomic and subatomic particles, we have examples where both energy and linear momentum are conserved.

Figure (22), shown previously in Chapter 6 as Figure (6-3), shows the track of a proton as it moves through a hydrogen bubble chamber. The incoming track ends when the proton strikes a hydrogen nucleus (another proton) that is part of the liquid hydrogen. Three dimensional stereoscopic photographs show that the two protons recoil from each other at an angle of 90°. Here we are looking at the behavior of matter on a subatomic scale where Newtonian mechanics does not apply, and we would like to find out what we can learn from this collision.

Does this photograph show us that momentum and energy are still conserved on the subatomic scale? To find out we will analyze the collision of larger objects like steel ball bearings, where momentum is conserved and energy nearly so, and see if the results explain what we see in Figure (22).



**Figure 22** An incoming proton collides with a proton at rest. The two protons recoil at right angles.

#### **Elastic Collisions**

We will start with the simplest elastic collision we can think of—a ball of mass is traveling at a velocity  $\vec{v}_i$ strikes an identical ball at rest, as shown in Figure (23). We will assume that the collision is straight on so that the two balls go off in the same direction at speeds  $v_1$ and  $v_2$  as shown. The idea is to apply conservation of momentum and energy in order to predict the final speeds  $v_1$  and  $v_2$ .

From conservation of momentum we get

 $mv_i = mv_1 + mv_2$  momentum (27)

Canceling the m's we have

 $v_i = v_1 + v_2$  (28)

From conservation of energy we have

$$\frac{1}{2}mv_{i}^{2} = \frac{1}{2}mv_{1}^{2} + \frac{1}{2}mv_{2}^{2} \quad \frac{energy}{conservation}$$
(29)

The 1/2 m's cancel and we have

$$\mathbf{v}_i^2 = \mathbf{v}_1^2 + \mathbf{v}_2^2 \tag{30}$$

If we square Equation (28) we get

$$\mathbf{v_i}^2 = \mathbf{v_1}^2 + 2\mathbf{v_1}\mathbf{v_2} + \mathbf{v_2}^2 \tag{31}$$

Comparing this with Equation (30) we see that

 $v_1 v_2 = 0$ 

I.e., either  $v_2 = 0$ , which means there was no collision at all, or  $v_1 = 0$  which means that ball 1 stops dead in its tracks and ball 2 goes on at the initial speed  $v_i$ .



**Figure 23** Collision of two identical steel balls. We want to calculate  $v_1$  and  $v_2$ .

That ball 1 stops dead in its tracks explains the common toy where two or more steel balls are suspended from a string as shown in the end view—Figure (24a). One of the balls is pulled back as shown in Figure (24b) and released. At the collision, it comes to rest and the second ball goes on. Then the second ball comes back, strikes the first one and stops, and the first ball goes back up. For good hard steel balls, this process goes on for a long time before we see motion decrease.

If energy were not conserved, if some were lost in the collision, then both balls would move forward after the collision. In the extreme case the balls would stick and move off together. This is a completely inelastic collision where the maximum energy is lost (consistent with conservation of momentum). If you perform this experiment and notice that the incoming ball really comes to rest, that is experimental proof that both energy and momentum were in fact conserved in the collision.



#### Figure 24

When the balls collide, the moving ball stops and the stationary ball moves on at the same speed.

In the next simplest example, suppose the two balls of Figure (23) collide but bounce off at an angle as shown in Figure (25), ball 1 coming off at a velocity  $\vec{v}_1$ , and ball 2 at a velocity  $\vec{v}_2$  as shown. In this case we have the same formula for conservation of energy, but conservation of momentum must now be written as the vector equation

$$m\vec{v}_i = m\vec{v}_1 + m\vec{v}_2$$
 momentum  
conservation (32)

Again the m's cancel and we are left with the following vector equation for the velocity vectors

$$\vec{v}_1 = \vec{v}_1 + \vec{v}_2$$
 (32a)

The equation if pictured in Figure (26).

Recall that energy conservation gave

$$v_i^2 = v_1^2 + v_2^2 \qquad \begin{array}{c} energy\\ conservation \end{array} \tag{30}$$

which is simply the *Pythagorean theorem* when applied to the triangle in Figure (26). Thus the incoming speed  $\vec{v}_i$  must be the hypotenuse and  $\vec{v}_1$  and  $\vec{v}_2$  the sides of a right triangle. Energy conservation requires that the two balls emerge from the collision at right angles.



**Figure 25** If the collision is not straight on, the balls come off at an angle.



**Figure 26** *Equation 30 requires that the velocity vectors form a right triangle.* 

To state this result another way, if two equal masses collide and one is originally at rest, they always emerge at right angles (or the incoming one stops). This is experimental evidence that both energy and momentum are conserved in the collision.

It is this way that we learn that energy and momentum are both conserved during the collision of two protons in a hydrogen bubble chamber. This is a remarkable result considering that we do not have to know anything about what kind of forces were involved in the collision.

If we have elastic collisions between objects of different masses, moving at each other with different speeds, as in Figure (27), we still have the conservation of momentum

$$m_1 \vec{v}_{1\,i} + m_2 \vec{v}_{2\,i} = m_1 \vec{v}_{1\,f} + m_2 \vec{v}_{2\,f}$$
(33)

and conservation of energy

$$\frac{1}{2}m_1v_{1\,i}^2 + \frac{1}{2}m_2v_{2\,i}^2 = \frac{1}{2}m_1v_{1\,f}^2 + \frac{1}{2}m_2v_{2\,f}^2$$
(34)

The only problem is that the algebra quickly becomes messy.

If you are in the business of working with collision problems, you will find it much easier to go to a coordinate system where the center of mass of the colliding particles is at rest. In this coordinate system the particles go in and come out symmetrically and the equations are easy to solve. Then you transform back to your original coordinate system to see what the particle velocities should be in the laboratory.



**Figure 27** *Arbitrary collision of two balls.* 

## **DISCOVERY OF THE ATOMIC NUCLEUS**

In his early experiments with radioactivity, Ernest Rutherford found that radioactive atoms emitted three kinds of radiation which, as we have mentioned, he called  $\alpha$  rays,  $\beta$  rays and  $\gamma$  rays. The  $\alpha$  rays turned out to be heavy positively charged particles, later identified as helium nuclei. The  $\beta$  rays were beams of negatively charged particles later determined to be electrons, and the neutral  $\gamma$  rays turned out to be high energy particles of light (photons).

In the early 1900s, before 1912, it was not clear how these particles were emitted or what the structure of the atom was. Since J. J. Thomson's experiments with electron beams in 1895, it was known that atoms contained electrons, and it was also known that complete atoms were electrically neutral and much more massive than an electron. Thus the atom had to have mass and positive charge in some form or other, but no one knew what form. By 1912 the plum pudding model was quite popular. This was a picture in which mass and positive charge was spread throughout the atom like the pudding, and the electrons were located at various points, like the plums. A rather vague picture at best.

In 1912 Rutherford and Hans Geiger began a series of experiments using beams of radioactive particles to probe the structure of matter. These experiments could begin after Geiger had developed a tube to detect radioactive particles. This device later became known as a *Geiger counter*, and is still used through the world to monitor radiation.

In the first set of experiments, a beam of  $\alpha$  particles were aimed at a gold foil. It was expected that some of the  $\alpha$  particles would be slightly deflected as they passed through the positive matter in the gold atoms, or came near electrons. To the utter amazement of both Rutherford and Geiger, some of the  $\alpha$  particles bounced straight back out of the gold foil, with essentially the same kinetic energy they had going in. We have seen from our analysis of the elastic collision of two equal mass particles, that the incoming particle stops and the struck particle continues on. Only if the mass of the struck particle is greater than the mass of the incoming particle, will the incoming particle bounce back. And only if the mass of the struck particle is much greater than the mass of the incoming particle will the incoming particle rebound with nearly the same energy that it had coming in. Thus Rutherford and Geiger's observation that some of the  $\alpha$  particles bounced right back out of the gold foil, indicated that they struck a solid object much more massive than an  $\alpha$  particle.

Most  $\alpha$  particles passed through the gold foil without much deflection, indicating that most of the volume of the gold foil was devoid of mass. The few collisions that did result in a recoil indicated that the mass in a gold foil was concentrated in incredibly small regions of space. A more detailed analysis showed that the scattering was caused by an electric force, thus they knew that both the mass and positive charge were located in a tiny region of the atom. In this way the atomic nucleus was discovered.

## NEUTRINOS

The discovery of the neutrino, or at least the prediction of its existence, is another important event in the history of physics that is related to the conservation of energy and linear momentum.

After Rutherford's discovery of the nucleus, it became clear that the high energy radioactive emissions,  $\alpha$ ,  $\beta$ , and  $\gamma$  rays, must be coming from the nucleus of the atom. Thus a study of these rays should give valuable information about the nature of the nucleus itself.

After a number of years of experimentation it was determined that whenever an  $\alpha$  particle or  $\gamma$  ray was emitted, the energy carried out by the  $\alpha$  particle or  $\gamma$  ray was precisely equal to the energy lost by the nucleus. But decays involving  $\beta$  particles were different. In studying  $\beta$  decays, one always got a spread of energies of the  $\beta$  particle. Sometimes the  $\beta$  particle carried out almost all the energy lost by the nucleus, and sometimes only the relatively small rest energy of the electron. By the late 1920s it was clear that energy was apparently not conserved in  $\beta$  decay reactions. Neils Bohr proposed that the law of conservation of energy had to be modified for nuclear reactions involving  $\beta$  decays. The new rule was that the final energy was always less than or equal to the initial energy.

In 1930 Wolfgang Pauli, one of the founders of quantum theory, objected to the idea that energy was not conserved in  $\beta$  decay events. Pauli noted that the conservation of energy, linear momentum, and angular momentum are all apparently violated at the same time in  $\beta$  decay. Either the entire structure of physical law was being violated, or there was another explanation. Pauli's other explanation was that the energy, the linear momentum, and the angular momentum were all being carried out at the same time by an unseen particle. If Pauli's particle existed, it would need the following properties: (1) it had to be electrically neutral because no electric charge was lost in  $\beta$  decays; (2) it would have to have a very small rest mass because the electron or  $\beta$  particle sometimes carried out almost all the available energy, leaving none for creating the new particle's rest mass; (3) the new particle had to have almost no interaction with matter, otherwise someone would have seen it.

Initially there was not much enthusiasm for Pauli's idea of an undetectable particle. At that point no one had seen an electrically neutral particle, and the fact that it did not interact with matter made it seem too speculative.

In 1932 the neutron was discovered which demonstrated that neutral particles did exist. Shortly after that Enrico Fermi developed a detailed theory of the weak interaction in which neutrinos played a significant role. (Fermi called Paul's particle the *neutrino*, or "little neutral one" to distinguish it from the more massive neutron.) Detailed verification of Fermi's weak interaction theory convinced the physics community that neutrinos should exist.

The neutrino, actually detected in 1956 and now commonly seen in numerous experiments, is a remarkable particle in that it is subject to only the weak and gravitational interactions. All other known particles are subject to the electric and nuclear forces. For example  $\pi$  mesons are subject to the nuclear force, and can travel only a short distance through matter before colliding with a nucleus and being stopped.

The muon discussed in the relativity chapter, does not feel the strong nuclear force and therefore can travel much farther through matter (hundreds of meters) before being stopped. Muons are electrically charged and therefore are stopped by the weaker electric interaction. Photons also interact through the electric interaction, and therefore have a limited range traveling through matter. (An X-ray is an example of a photon passing through matter.) The weak interaction is so weak compared to the nuclear or electric force, that a particle like the neutrino which feels only the weak interaction force, can travel incredible distances through matter before being stopped. To have a good chance of stopping a single neutrino with a stack of lead, one would need a pile of lead, light years thick.

This does not mean that neutrinos are impossible to detect. Instead of using a detector light years across to detect one neutrino, one can use a source that produces an incredible number of neutrinos and use a reasonable sized detector so that one has some chance of stopping a few neutrinos. In 1956 Cowan and Rines placed a tank car full of carbon tetrachloride cleaning fluid in front of a nuclear reactor that was estimated to emit about 10<sup>15</sup> neutrinos per square centimeter per second. They observed that about two chlorine atoms per month in the tank car of carbon tetrachloride were converted by neutrino interactions into argon atoms which were counted individually.

In modern experiments carried out using high energy particle accelerators, neutrino reactions are routinely seen. The reason that more neutrinos are detected in these experiments is that the weak interaction becomes less weak as the energy of the particles involved increases. The high energy accelerators produce neutrinos with great enough energy that they are not too difficult to detect. As a result, the neutrino interactions has become an important research tool in the study of the basic interactions of matter. Neutrinos make a particularly clean tool for these studies because they have no nuclear or electric interactions. Neutrino experiments are not contaminated by effects of the nuclear and electric forces.

#### **Neutrino Astronomy**

An exciting development involving neutrinos is the birth of neutrino astronomy. In the fusion reaction that powers our sun, where four hydrogen nuclei (protons) end up as a helium 4 nucleus, the weak interaction and the  $\beta$  decay process comes into play in the conversion of two of the protons into the neutrons of the helium nucleus. Thus the emission of neutrinos must accompany the fusion reaction, and the neutrinos themselves must carry off a significant amount of the energy liberated by the fusion reaction.

In a star like the sun, the fusion reaction takes place down in the core of the star where the temperatures are highest. Any light emitted by the fusion reaction should take the order of about 10,000 years to work its way out. Thus if the fusion reaction in the sun were shut off today, it would be roughly 10,000 years before the sun dimmed.

Neutrinos, however, escape from the core of the sun without delay. If the fusion reaction stopped and we were monitoring the neutrinos from the sun, we would know about it within 8 minutes. As a result there is considerable incentive to observe the solar neutrinos, for that gives us a picture of what is happening in the sun's core now.

To study solar neutrinos, and do other experiments like look for decay of the proton, several large neutrino detectors have been set up around the world. Solar neutrinos have been monitored fairly carefully for over a decade, and there is an unexplained, perhaps disturbing result. Only about one third as many neutrinos are being emitted by the sun as we expect from what we think the fusion reaction should produce. Perhaps we are not detecting all we should, but the detectors are getting better and the number remains at 1/3. This is one of the major puzzles of astronomy. That neutrino astronomy is really here was dramatically illustrated with the supernova explosion of 1987. On the average, supernovas occur about once per century per galaxy. Kepler saw the last supernova explosion in our galaxy 400 years ago. In 1987 a graduate student spotted the sudden appearance of a bright star in the large Magellanic cloud, a close small neighboring galaxy. This was the first supernova explosion in the local region of our galaxy in 400 years.

In a supernova explosion, huge quantities of neutrinos should be emitted. In fact a fair fraction of the energy of the explosion should be carried out by neutrinos.

Theoretical models of supernova explosions suggest that light should take about three hours to work its way out through the expanding envelope of gas before it starts its trek through space at the speed c. Neutrinos, on the other hand, should escape without being slowed down, and have a three hour head start on the light. If neutrinos have no rest mass, and therefore travel at the speed of light, they should have reached the earth about three hours before the light. Two of the major neutrino detectors, one in the US and one in a tunnel in the Alps, detected significant pulses of neutrinos about three hours before the flare-up of the star was seen. (This was determined by a later analysis of the neutrino data.) That event marks the birth of neutrino astronomy on a galactic scale.

# Chapter 12 Rotational Motion

Our discussion of rotational motion begins with a review of the measurement of angles using the concept of radians. We will refer to an angle measured in radians as an **angular distance**. If we are discussing an object that is rotating, we will describe the rotation in terms of the increase in angular distance, namely an **angular velocity**. And if the speed of rotation is changing, we will describe the change in terms of an **angular acceleration**.

In Chapter 7, linear momentum and angular momentum were treated as distinctly separate topics. The main point of this chapter is to develop a close analogy between the two concepts. The linear momentum of an object is its mass m times its linear velocity  $\vec{v}$ . We will see that angular momentum can be expressed as an angular mass times an angular velocity. (Angular mass is more commonly known as moment of inertia). Then, using the formalism of the vector cross product (mentioned in Chapter 2), we will see that angular momentum can be treated as a vector quantity, which explains the bicycle wheel experiments we discussed in Chapter 7.

The fundamental concept of Newtonian mechanics is that the total force  $\vec{F}$  acting on an object is equal to the time rate of change of the object's linear momentum;  $\vec{F} = d\vec{p}/dt$ . Using the vector cross product formalism, we will obtain a complete angular analogy to this equation. We will find that a quantity we call an angular force is equal to the time rate of change of angular momentum. (The angular force is more commonly known as torque).

The angular analogy to Newton's second law looks a bit peculiar at first. It involves lever arms and vectors that point in funny directions. After some demonstrations to show that the equation appears to give reasonable results, we apply the equation to predict the motion of a gyroscope. The prediction appears to be absurd, but we find that that is the way a gyroscope behaves.

Our focus in this chapter is on angular momentum because that concept will play such an important role in our later discussions of atomic physics and electrons and nuclear magnetic resonance. There are other important and interesting topics such as rotational kinetic energy and the calculation of moments of inertia which we discuss in more detail in the appendix. These topics are not difficult and lead to some good lecture demonstrations and laboratory experiments. We put them in an appendix because they do not play the essential role that angular momentum does in our later discussions.

## **RADIAN MEASURE**

From the point of view of doing calculations, it is more convenient to measure an angle in radians than the more familiar degrees. In radian measure the angle  $\theta$ shown in Figure (1) is the ratio of the arc length s to the radius r of the circle

$$\theta \equiv \frac{S}{r}$$
 radians (1)

Since s and r are both distances, the ratio s/r is a dimensionless quantity. However we will find it convenient for the angular analogy to keep the name radians as if it were the actual dimension of the angle. For example we will measure angular velocities in *radians per second*, which is analogous to linear velocities measured in meters per second.

Since the circumference of a circle is  $2\pi r$ , the number of radians in a complete circle is

$$\Theta \begin{pmatrix} \text{complete} \\ \text{circle} \end{pmatrix} = \frac{2\pi r}{r} = 2\pi$$

In discussing rotation, we will often refer to going around one complete time as one complete *cycle*. In one cycle, the angle  $\theta$  increases by  $2\pi$ . Thus  $2\pi$  is the number of *radians per cycle*. We will find it convenient to assign these dimensions to the number  $2\pi$ :



The angle  $\theta$  in radians is defined as the ratio of the arc length s to the radius r:  $\theta = s/r$ .

To relate radians to degrees, we use the fact that there are 360 degrees/cycle and dimensional analysis to find the number of degrees/radian

$$360 \frac{\text{degrees}}{\text{cycle}} \times \frac{1}{2\pi} \frac{\text{cycles}}{\text{radian}} = \frac{360}{2\pi} \frac{\text{degrees}}{\text{radian}} = 57.3 \frac{\text{degrees}}{\text{radian}}$$
(3)

Fifty seven degrees is a fairly awkward unit angle for

purposes of drafting and navigation; no one in his or her right mind would mark acompass in radians. However, in working with the dynamics of rotational mo-



tion, radian measure is the only reasonable choice.

## Angular Velocity

The typical measure of angular velocity you may be familiar with is revolutions per minute (RPM). The tachometer in a sports car is calibrated in RPM; a typical sports car engine gives its maximum power around 5000 RPM. Engine manufacturers in Europe are beginning to change over to revolutions per second (RPS), but somehow revving an engine up to 83 RPS doesn't sound as impressive as 5000 RPM. (Tachometers will probably be calibrated in RPM for a while.) In physics texts, angular velocity is measured in radians per second. Since there are  $2\pi$  radians/cycle, 83 revolutions or cycles per second corresponds to  $2\pi \times 83 = 524$  radians/second. Few people would know what you were talking about if you said that you should shift gears when the engine got up to 524 radians per second.

#### **Exercise 1**

What is the angular velocity, in radians per second, of the hour hand on a clock?

Angular Analogy

At this point we have a complete analogy between the rotation of a motor shaft and one dimensional linear motion. This analogy becomes clear when we write out the definitions of position, velocity, and acceleration:

	Linear motion	Angular motion
Distance	X meters	$\theta$ radians
Velocity	$v = \frac{dx}{dt} \frac{meters}{second}$	$\omega = \frac{\mathrm{d}\theta}{\mathrm{d}t} \frac{\mathrm{radians}}{\mathrm{second}}$
Acceleration	$a = \frac{dv}{dt} \frac{meters}{second^2}$	$\alpha = \frac{d\omega \text{ radians}}{dt \text{ second}^2}$
	$=\frac{d^2x}{dt^2}$	$=\frac{d^2\theta}{dt^2} \qquad (7)$

As far as these equations go, the analogy is precise. Therefore any formulas that we derived for linear motion in one dimension must also apply to angular motion. In particular the constant acceleration formulas, derived in Chapter 3, must apply. If the linear and angular accelerations a and  $\alpha$  are constant, then we get

#### **Constant Acceleration Formulas**

Linear motion (a = const)	Angular motion ( $\alpha = const$ )	
$x = v_0 t + \frac{1}{2}at^2$	$\theta = \omega_0 t + \frac{1}{2} \alpha t^2$	(8)
$v = v_0 + at$	$\omega = \omega_0 + \alpha t$	(9)

#### Exercise 2

An electric motor, that turns at 3600 rpm (revolutions per minute) gets up to speed in 1/2 second. Assume that the angular acceleration  $\alpha$  was constant while the motor was getting up to speed.

a) What was  $\alpha$  (in radians/sec<sup>2</sup>)?

b) How many radians, and how many complete cycles, did the shaft turn while getting up to speed?

Our formal definition of angular velocity is the time rate of change of an angle. We almost always use the Greek letter  $\omega$  (omega) to designate angular velocity

$$\frac{angular}{velocity} \omega \equiv \frac{d\theta}{dt} \frac{radians}{second}$$

When thinking of angular velocity  $\omega$  picture a line marked on the end of a rotating shaft. The angle  $\theta$  is the angle that the line makes with the horizontal as shown in Figure (2). As the shaft rotates, the angle  $\theta(t)$  increases with time, increasing by  $2\pi$  every time the shaft goes all the way around.

(4)

#### Angular Acceleration

When we start a motor, the angular velocity of the shaft starts at  $\omega = 0$  and increases until the motor gets up to its normal speed. During this start-up,  $\omega(t)$  changes with time, and we have an angular acceleration  $\alpha$ defined by

$$\frac{angular}{acceleration} \alpha \equiv \frac{d\omega}{dt} \frac{radians}{second^2}$$
(5)

The angular acceleration  $\alpha$  has the dimensions of radians/sec<sup>2</sup> since the derivative gives us another factor of time in the denominator. Combining Equation 4 and 5 relates  $\alpha$  to  $\theta$  by



End of a shaft rotating at an angular velocity  $\omega$ .

## Tangential Distance, Velocity and Acceleration

So far we have used the model of a rotating shaft to illustrate the concepts of angular distance, velocity and acceleration. We now wish to shift the focus of our discussion to the dynamics of a particle traveling along a circular path. For this we will use the model of a small mass m on the end of a massless stick of length r shown in Figure (3). The other end of the stick is attached to and is free to rotate about a fixed axis at the origin of our coordinate system. The presence of the stick ensures that the mass m travels only along a circular path of radius r. The quantity  $\theta(t)$  is the angular distance travelled and  $\omega(t)$  the angular velocity of the particle.

When we are discussing the motion of a particle in a circular orbit, we often want to know how far the particle has travelled, or how fast it is moving. The distance s along the path (we could call the tangential distance) travelled is given by Equation 1 as





The speed of the particle along the path, which we can call the tangential speed  $v_t$ , is the time derivative of the tangential distance s(t)

$$v_t = \frac{ds(t)}{dt} = \frac{d}{dt} [r\theta(t)] = r \frac{d\theta(t)}{dt}$$

where r comes outside the derivative since it is constant. Since  $d\theta(t)/dt$  is the angular velocity  $\omega$ , we get

$$v_{t} = r\omega \qquad tangential \\ velocity \qquad (11)$$

The tangential acceleration  $a_t$ , the acceleration of the particle along its path, is the time derivative of the tangential velocity

$$a_{t} = \frac{dv_{t}(t)}{dt} = \frac{d[r\omega(t)]}{dt} = r\frac{d\omega(t)}{dt} = r\alpha$$

$$a_{t} = r\alpha \qquad \underset{acceleration}{\overset{tangential}{acceleration}} \qquad (12)$$

where again we took the constant r outside the derivative, and used  $\alpha = d\omega/dt$ .



#### Figure 4

Particle moving at a constant speed in a circle of radius r accelerates toward the center of the circle with an acceleration of magnitude  $a_r = v^2/r$ .

#### **Radial** Acceleration

If the angular velocity  $\omega$  is constant, if we have a particle traveling at constant speed in a circle, then  $\alpha = d\omega/dt = 0$  and there is no tangential acceleration  $a_t$ . However, we have known from almost the beginning of the course that a particle traveling at constant speed v in a circle of radius r has an acceleration directed toward the center of the circle, of magnitude  $v^2/r$ , as shown in Figure (4). We will now call this center directed acceleration the *radial acceleration*  $a_r$ 

$$a_{r} = \frac{v_{t}^{2}}{r} \quad \begin{array}{c} radial \\ acceleration \end{array}$$
(13)

#### **Exercise 3**

Express the radial acceleration  $a_r$  in terms of the orbital radius r and the particle's angular velocity  $\omega$ .

If a particle is traveling in a circular orbit, but its speed  $v_t$  is not constant, then it has both a radial acceleration  $a_r = v_t^2/r$ , and a tangential acceleration  $a_t = r\alpha$ . The radial acceleration is always directed toward the center of the circle and always has a magnitude  $v^2/r$ . The tangential acceleration, if it exists, is tangential to the circle, pointing forward (counterclockwise) if  $\alpha$  is positive and backward if  $\alpha$  is negative. These accelerations are shown in Figure (5).



**Figure 5** *Motion with radial and tangential acceleration*.

#### **Bicycle Wheel**

For much of the remainder of the chapter, we will use a bicycle wheel, often weighted with wire wound around the rim, to illustrate various phenomena of rotational motion. Conceptually we can think of the bicycle wheel as a collection of masses on the ends of massless rods as shown in Figure (6). The massless rods form the spokes of the wheel, and we can think of the masses m as fusing together to form the wheel. When forming a wheel, all the masses have the same radius r, same angular velocity  $\omega$  and same angular acceleration  $\alpha$ . If we choose one point on the wheel from which to measure the angular distance  $\theta$ , then as far as angular motion is concerned, it does not make any difference whether we are discussing the mass on the end of a rod shown in Figure (3) or the bicycle wheel shown in Figure (6). Which model we use depends upon which provides a clearer insight into the phenomena being discussed.



Bicycle wheel as a collection of masses on the end of massless rods.

Figure 6

## ANGULAR MOMENTUM

In Chapter 7, we defined the angular momentum  $\ell$  of a mass m traveling at a speed v in a circle of radius r as

$$l = mvr$$

As we saw, in Figure (7-9) reproduced here, the quantity  $\ell = mvr$  did not change when we had a ball moving in a circle on the end of a string, and we pulled in on the string. The radius of the circle decreased, but the speed increased to keep the product (vr) constant.



Figure 7–9 Ball on the end of a string, swinging in a circle.

This was our introduction to the concept of the conservation of angular momentum.

After that, we went on to consider some rather interesting experiments where we held a rotating bicycle wheel while standing on a freely turning platform. We found that these experiments could be explained qualitatively if we thought of the angular momentum of the bicycle wheel as being a vector quantity  $\vec{\ell}$  which pointed along the axis of the wheel, as shown in Figure (7-15) reproduced below. What we will do now is develop the formalism which treats angular momentum as a vector.



#### Figure 7–15

When the bicycle wheel is turned over and its angular momentum points down, the person starts rotating with twice as much angular momentum, pointing up.

## Angular Momentum of a Bicycle Wheel

We will begin our discussion of the angular momentum of a bicycle wheel using the picture of a bicycle wheel shown in Figure (6), i.e., a collection of balls on the end of massless rods or spokes. If the wheel is rotating with an angular velocity  $\omega$ , then each ball has a tangential velocity v<sub>t</sub> given by Equation 11a

$$v_t = r\omega$$
 (1 hepeated)

If the i-th ball in the wheel (identified in Figure 7) has a mass  $m_i$ , then its angular momentum  $\ell_i$  will be given by

$$\ell_{i} = m_{i}v_{t}r = m_{i}(r\omega)r$$

$$\ell_{i} = (m_{i}r^{2})\omega \qquad (14)$$

Assuming that the total angular momentum L of the bicycle wheel is the sum of the angular momenta of each ball (we will discuss this assumption in more detail shortly) we get

$$L = \sum_{i} \ell_{i} = \sum_{i} \left( m_{i} r^{2} \omega \right)$$
(15)

Since each mass  $m_i$  is at the same radius r and is traveling with the same angular velocity  $\omega$ , we get

$$L = \left(\sum_{i} m_{i}\right) r^{2} \omega$$

Noting that  $M = \left(\sum_{i} m_{i}\right)$  is the total mass of the bicycle wheel, we get



Figure 7

The angular momentum of the i'th ball is  $m_i v_t r_i$ .

## Angular Velocity as a Vector

To explain the bicycle wheel experiment discussed in Chapter 7, we assumed that the angular momentum  $\vec{L}$  was a vector pointing along the axis of the wheel as shown in Figure (8a). We can obtain this vector concept of angular momentum by first defining a vector angular velocity  $\vec{\omega}$  as shown in Figure (8b). We will say that if a wheel is rotating with an angular velocity  $\omega_{rad/sec}$ , the vector  $\vec{\omega}$  has a magnitude of  $\omega_{rad/sec}$ , and points along the axis of rotation as shown in Figure (8b). Since the axis has two directions, we use a right hand convention to select among them. Curl the fingers of your right hand in the direction of the direction of the rotation, and the thumb of your right hand will point in the direction of the vector  $\vec{\omega}$ .

#### Angular Momentum as a Vector

Since the vector  $\vec{\omega}$  points in the direction we want the angular momentum vector  $\vec{L}$  to point, we can obtain a vector formula for  $\vec{L}$  by simply replacing  $\omega$  by  $\vec{\omega}$  in Equation 16 for the angular momentum of the bicycle wheel



## ANGULAR MASS OR MOMENT OF INERTIA

Equation 17 expresses the angular momentum  $\vec{L}$  of a bicycle wheel as a numerical quantity  $(Mr^2)$  times the vector angular velocity  $\vec{\omega}$ . This is not very different from linear momentum  $\vec{p}$  which is the mass (M) times the linear velocity vector  $\vec{v}$ 

$$\vec{p} = M\vec{v} \tag{18}$$

We obtain an analogy between linear and angular momentum if we call the quantity  $(Mr^2)$  the *angular mass* of the bicycle wheel. Designating the angular mass by the letter I, we get

$$\vec{L} = I\vec{\omega} \tag{19}$$



The quantity I is usually called *moment of inertia* rather than angular mass, but angular mass provides a better description of what we are dealing with. We will use either name, depending upon which seems more appropriate.

## **Calculating Moments of Inertia**

Equation 20 is not the most general formula for calculating moments of inertia. The bicycle wheel is special in that all the mass is essentially out at a single radius r. If, instead, we had a solid wheel where the mass was spread out over different radii, we would have to conceptually break the wheel into a number of separate rim-like wheels of radii  $r_i$  and mass  $m_i$ , calculate the moment of inertia of each rim, and add the results together to get the total moment of inertia.

In Appendix A we have relatively complete discussion of how to calculate moments of inertia, and how moment of inertia is related to rotational kinetic energy. There you will see that rotational kinetic energy is  $1/2 I\omega^2$ , which is analogous to the linear kinetic energy  $1/2 Mv^2$ . This material is placed in an appendix, not because it is difficult, but because we do not wish to digress from our discussion of the analogy between linear and angular momentum. At this point, one example and one exercise should be a sufficient introduction to the concept of moment of inertia.

#### Example 1

Calculate the moment of inertia, about its axis, of a cylinder of mass M and outside radius R. Assume that the cylinder has uniform density.

Solution: We conceptually break the cylinder into a series of concentric cylinders of radius r and thickness dr as shown in Figure (9). Each hollow cylinder has a mass given by

$$dm = M \times \frac{\text{end area of hollow cylinder}}{\text{total end area}}$$
$$= M \times \frac{2\pi r \, dr}{\pi R^2} = M \times \frac{2r \, dr}{R^2}$$

(21)

Since all the mass in the hollow cylinder is out at a radius r, just as it is for a bicycle wheel, the hollow cylinder has a moment of inertia dI given by

$$dI = (dm) \times r^{2}$$
$$= M \times \frac{2r dr}{R^{2}} \times r^{2} = \frac{2Mr^{3} dr}{R^{2}} \qquad (22)$$

The total moment of inertia of the cylinder is the sum of the moments of inertia of all the hollow cylinders. This addition is done by integrating the formula for dI from r = 0 out to r = R.



**Figure 9** Calculating the moment of inertia of a cylinder about its axis of rotation.

Two points are made in Example 1, The first is that calculating the moment of inertia of an object usually requires an integration, because different parts of the object are out at different distances r from the axis of rotation. Secondly we see that the moment of inertia of a solid cylinder is less than the moment of inertia of a bicycle wheel of the same mass and outer radius  $(1/2 MR^2$  for the cylinder versus  $MR^2$  for the bicycle wheel). This is because all the mass of the bicycle wheel is out at the maximum radius R, while most of the mass of the solid cylinder is in at smaller radii.

A considerable amount of time can be spent discussing the calculation of moments of inertia of various shaped objects. Rather than do that here, we will simply present a table of the moments of inertia of common objects of mass M and outer radius R, about an axis that passes through the center.

Object	Moment of Inertia
cylindrical shell	$1 \mathrm{MR}^2$
solid cylinder	$1/2 { m MR}^2$
spherical shell	2/3 MR <sup>2</sup>
solid sphere	$2/5 \mathrm{MR}^2$

#### **Exercise 4**

As shown in Figure (10) we have a thick-walled hollow brass cylinder of mass M, with an inner radius  $R_i$  and outer radius  $R_0$ . Calculate its moment of inertia about its axis of symmetry. Check your answer for the case  $R_i = 0$  (a solid cylinder) and for  $R_i = R_0$  (which corresponds to the bicycle wheel).



Thick-walled hollow cylinder.

Figure 10

The idea of having the angular velocity  $\vec{\omega}$  being a vector pointing along the axis of rotation gave us a nice analogy between linear momentum  $\vec{p} = M\vec{v}$  and angular momentum  $\vec{L} = I\vec{\omega}$ . But to obtain the dynamical equation for angular momentum, the one analogous to Newton's second law for linear momentum, we need the mathematical formalism of the vector cross product defined back in Chapter 2. Since we have not used the vector cross product before now, we will briefly review the topic here.

If we have two vectors  $\vec{A}$  and  $\vec{B}$  like those shown in Figure (11), the vector cross product  $\vec{A} \times \vec{B}$  is defined to have a magnitude

$$\left|\vec{\mathbf{A}} \times \vec{\mathbf{B}}\right| = \left|\vec{\mathbf{A}}\right| \left|\vec{\mathbf{B}}\right| \sin \theta \tag{24}$$

where  $|\vec{A}|$  and  $|\vec{B}|$  are the magnitudes of the vectors  $\vec{A}$  and  $\vec{B}$ , and  $\theta$  is the small angle between them. Note that when the vectors are parallel, sin  $\theta = 0$  and the cross





product is zero. The cross product is a maximum when the vectors are perpendicular. This is just the opposite from the scalar dot product which is a maximum when the vectors are parallel and zero when perpendicular. Conceptually you can think of the dot product as measuring parallelism while the cross product measures perpendicularity.

The other major difference between the dot and cross product is that with the dot product we end up with a number (a scalar), while with the cross product, we end up with a vector. The direction of  $\vec{A} \times \vec{B}$  is the most peculiar feature of the cross product; it is *perpendicular* to the plane defined by the vectors  $\vec{A}$  and  $\vec{B}$ . If we draw  $\vec{A}$  and  $\vec{B}$  on a sheet of paper as we did in Figure (11), then the directions perpendicular to both  $\vec{A}$  and  $\vec{B}$  are either up out of the paper or down into the paper. To decide which of these two directions to choose, we use the following right hand rule. (This is an arbitrary convention, but if you use it consistently in all of your calculations, everything works out OK). To find the direction of the vector  $\vec{A} \times \vec{B}$ , point the fingers of your right hand in the direction of the first vector in the product (namely  $\vec{A}$ ). Then, without breaking your knuckles, curl the finger of your right hand toward the second vector  $\vec{B}$ . Curl them in the direction of the small angle  $\theta$ . If you do this correctly, the thumb of your right hand will point in the direction of the cross product  $\vec{A} \times \vec{B}$ . Applying this to the vectors in Figure (11), we find that the vector  $\mathbf{A} \times \mathbf{B}$ points up out of the paper as shown in Figure (12).

#### **Exercise 5**

(a) Follow the steps we just mentioned to show that  $\vec{A} \times \vec{B}$  from Figure (11) does point up out of the paper.

(b) Show that the vector  $\vec{B} \times \vec{A}$  points down into the paper.



#### Figure 12

Right hand rule for vector cross product  $\vec{A} \times \vec{B}$ . Point the fingers of your right hand in the direction of the first vector  $\vec{A}$  and then curl them in the direction of the second vector  $\vec{B}$  (without breaking your knuckles). Your thumb will then point in the direction of the cross product  $\vec{A} \times \vec{B}$ .

Right Hand Rule for Cross Products u did the exercise (5b) correctly, you found that  $\vec{B} \times \vec{A}$  points in the opposite direction from the vector  $\vec{A} \times \vec{B}$ . In all previous examples of multiplication you have likely to have encountered, the order in which you did the multiplication made no difference. For example, both  $3 \times 5$  and  $5 \times 3$  give the same answer 15. But now we find that  $\vec{A} \times \vec{B} = -\vec{B} \times \vec{A}$  and the order of the multiplication does make a difference. Mathematicians say that cross product multiplication *does not* commute.

> There is one other special feature of the cross product worth noting. If  $\vec{A}$  and  $\vec{B}$  are parallel, or anti parallel. then they do not define a unique plane and there is no unique direction perpendicular to both of them. Various possibilities are indicated in Figure (13). But when the vectors are parallel or anti parallel,  $\sin \theta = 0$  and the cross product is zero. The special case where the cross product does not have a unique direction is when the cross product has zero magnitude with the result that the lack of uniqueness does not cause a problem.





If the vectors  $\vec{A}$  and  $\vec{B}$  are either parallel or antiparallel, then as shown above, there is a whole plane of vectors perpendicular to both  $\vec{A}$  and  $\vec{B}$ .

## CROSS PRODUCT DEFINITION OF ANGULAR MOMENTUM

Let us now see how we can use the idea of a vector cross product to obtain a definition of angular momentum vectors. To explain the bicycle wheel experiments, we wanted the angular momentum to point along the axis of the wheel as shown in Figure (14a). Since there are two directions along the axis, we have arbitrarily chosen the direction defined by the right hand convention shown. (Curl the fingers of your right hand in the direction of the rotation and your thumb will point in the direction of  $\vec{L}$ ).



In Figure (14b) we went to the masses and spoke model of the bicycle wheel, and selected one particular mass which we called  $m_i$ . This mass is located at a coordinate vector  $\vec{r}_i$  from the center of the wheel, and is traveling with a velocity  $\vec{v}_i$ . According to our definition of angular momentum in Chapter 7, using the formula  $\ell = mvr$ , the ball's angular momentum should be

 $\ell = m_i r_i v_i$ 

(25)

What we want to do now is to turn this definition of angular momentum into a vector that points down the axis of the wheel. This we can do with the vector cross product of  $\vec{r}_i$  and  $\vec{v}_i$ . We will try the definition of the vector  $\vec{\ell}_i$  as



#### **Exercise 6**

a) Look at Figure (14c) showing the vectors  $\vec{r}_i$  (which point into the paper) and  $\vec{v}_i$ . Point the fingers of your right hand in the direction of  $\vec{r}_i$  and then curl them toward the vector  $\vec{v}_i$ . Does your thumb point in the direction of the vector  $\vec{\ell}_i$  shown? (If it does not, you have peculiar knuckle joints or are not following instructions).



**Figure 14c** The three vectors  $\vec{r}_i, \vec{v}_i$  and  $\dot{\ell}_i$ 

b) Choose any other mass that forms the bicycle wheel shown in Figure (14b). Call that the mass  $m_i$ . Show that the vector  $\vec{\ell}_i = m_i (\vec{r}_i \times \vec{v}_i)$  also points down the axis, parallel to  $\vec{\ell}_i$ . Try this for several different masses, say one at the top, one at the front, and one at the bottom of the wheel.

If you did Exercise 6 correctly, you found that all the angular momentum vectors  $\vec{\ell}_i = m_i (\vec{r}_i \times \vec{v}_i)$  were parallel to each other, all pointing down the axis of the wheel. We will define the total angular momentum of the wheel as the vector sum of the individual angular momentum vectors  $\vec{\ell}_i$ 

$$\vec{L} \begin{pmatrix} total \ angular \\ momentum \\ of \ wheel \end{pmatrix} = \sum_{i} \vec{\ell}_{i} = \sum_{i} m_{i} \vec{r}_{i} \times \vec{v}_{i}$$
(26)

It is easy to add the vectors  $\vec{\ell}_i$  because they all point in the same direction, as shown in Figure (15). Thus we can add their magnitudes numerically. (It is just the numerical sum we did back in Equation 15).

**Figure 15** Since all the angular momentum vectors  $\dot{\ell}_i$  point in the same direction, we can add them up numerically.



To do the sum starting from Equation (26) we note that for each mass  $m_i$ , the vectors  $\vec{r}_i$  and  $\vec{v}_i$  are perpendicular, thus

$$|\vec{\mathbf{r}}_{i} \times \vec{\mathbf{v}}_{i}| = \mathbf{r} \mathbf{v} \sin \theta$$
  
=  $\mathbf{r} \mathbf{v} \quad (for \theta = 90^{\circ})$ 

Then note that for a rotating wheel, the speed v of the rim is related to the angular velocity  $\omega$  by

$$v = r\omega$$
 (1 hepeated)

so that

$$\left| \vec{r}_{i} \times \vec{v}_{i} \right| = rv = r(r\omega) = r^{2}\omega$$
 (27)

Finally note that the vector  $\vec{\omega}$  points in the same direction as  $\vec{r}_i \times \vec{v}_i$ , so that Equation 27 can be written as the vector equation

$$\vec{r}_i \times \vec{v}_i = r^2 \vec{\omega} \qquad for all \\ mass \ m_i$$
 (28)

Using Equation 28 in Equation 26 gives

$$\vec{L} = \sum_{i} \vec{\ell}_{i} = \sum_{i} m_{i} \vec{r}_{i} \times \vec{v}_{i}$$

$$= \left(\sum_{i} m_{i}\right) r^{2} \vec{\omega}$$

$$= Mr^{2} \vec{\omega}$$

$$\vec{L} = Mr^{2} \vec{\omega}$$
angular momentum
of a rotating
bicycle wheel
(29)

where M is the sum of the individual masses  $m_i$ . Equation 29 is the desired vector version of our original Equation 16.

The important point to get from the above discussion is that by using the vector cross product definition of angular momentum  $\vec{\ell}_i = m_i \vec{r}_i \times \vec{v}_i$ , all the  $\vec{\ell}_i$  for each mass in the wheel pointed down the axis of the wheel, and we could thus calculate the total angular momentum by numerically adding up the individual  $\vec{\ell}_i$ .

## The $\vec{r}\times\vec{p}$ Definition of Angular Momentum

A slight rewriting of our definition of angular momentum, Equation 25, gives us a more compact, easily remembered result. Noting that the linear momentum  $\vec{p}$  of a particle is  $\vec{p} = m\vec{v}$ , then a particle's angular momentum  $\vec{\ell}$  can be written

$$\vec{\ell} = \mathbf{m}\vec{\mathbf{r}} \times \vec{\mathbf{v}}$$
$$= \vec{\mathbf{r}} \times (\mathbf{m}\vec{\mathbf{v}})$$
$$\vec{\ell} = \vec{\mathbf{r}} \times \vec{\mathbf{p}}$$
(30)

In Chapter 7, we saw that the magnitude of the angular momentum  $\ell$  if a particle was given by the formula

$$\ell = r_{\perp} p \tag{7-15}$$

where  $r_{\perp}$  was the lever arm or perpendicular distance from the path of the particle to the point O about which we were measuring the angular momentum. This was illustrated in Figure (7-10) (reproduced here), where a ball of momentum  $\vec{p}$ , passing by an axis O, is caught by a hook and starts rotating in a circle.





After the ball is caught it is traveling in a circle with an angular momentum  $\ell = r_{\perp}mv = r_{\perp}p$ . By defining the angular momentum as  $r_{\perp}p$  even before the ball was caught, we could say that the ball had the same angular momentum  $r_{\perp}p$  before it was caught by the hook as it did afterward; that the angular momentum was unchanged when the ball was grabbed by the hook.

The idea that the angular momentum is the linear momentum times the perpendicular lever arm  $r_{\perp}$  follows automatically from the cross product definition of angular momentum  $\vec{\ell} = \vec{r} \times \vec{p}$ . To see this, consider a ball with momentum  $\vec{p}$  moving past an axis O as shown in Figure (16a). At the instant of time shown, the ball is located at a coordinate vector  $\vec{r}$  from the axis. The angle between the vectors  $\vec{r}$  and  $\vec{p}$  is the angle  $\theta$  shown in Figure (16b). The vector cross product  $\vec{r} \times \vec{p}$  is given

$$\left|\vec{\ell}\right| = \left|\vec{r} \times \vec{p}\right| = rp \sin \theta$$
 (31)

However we note that the lever arm or perpendicular distance  $r_{\perp}$  is given from Figure (16a)

 $\mathbf{r}_{\perp} = \mathbf{r}\sin\theta \tag{32}$ 

Combining Equations 31 and 32 gives

$$\left|\vec{\ell}\right| = \left|\vec{r} \times \vec{p}\right| = (r \sin\theta)p = r_{\perp}p \tag{33}$$

which is the result we used back in Chapter 7.

**Figure 16a** The coordinate vector  $\vec{r}$  and the lever arm  $r_{\perp}$  are related by  $r_{\perp} = r \sin \theta$ .



**Figure 16b** The angle between  $\vec{r}$  and  $\vec{p}$  is  $\theta$ .



#### Exercise 7

Using the vectors  $\vec{r}$  and  $\vec{p}$  in Figure (16), does the vector  $\vec{l} = \vec{r} \times \vec{p}$  point up out of the paper or down into the paper?

The intuitive point you should get from this discussion is that the magnitude of the vector cross product  $\vec{r} \times \vec{p}$ is equal to the magnitude of p times the perpendicular lever arm  $r_{\perp}$ . We will shortly encounter the cross product  $\vec{r} \times \vec{F}$  where  $\vec{F}$  is a force vector. We will immediately know that the magnitude of  $\vec{r} \times \vec{F}$  is  $r_{\perp}F$ where again  $r_{\perp}$  is a perpendicular lever arm.

## ANGULAR ANALOGY TO NEWTON'S SECOND LAW

We now have the mathematical machinery we need to formulate a complete angular analogy to Newton's second law. We do this by noting that to go from linear momentum  $\vec{p}$  to angular momentum  $\vec{\ell}$ , we took the cross product with the coordinate vector  $\vec{r}$ 

$$\ell = \vec{r} \times \vec{p}$$
 (Oppended)

The origin of the coordinate vector  $\vec{r}$  is the point about which we wish to calculate the angular momentum.

To obtain a dynamical equation for angular momentum  $\vec{\ell}$ , we start with Newton's second law which is a dynamical equation for linear momentum  $\vec{p}$ 

$$\vec{F} = \frac{d\vec{p}}{dt}$$
(11-16)

where  $\vec{F}$  is the vector sum of the forces acting on the particle.

With one mathematical trick, we can reexpress Newton's second law in terms of angular momentum. The mathematical trick involves evaluating the expression

$$\frac{\mathrm{d}}{\mathrm{d}t}(\vec{\mathbf{r}}\times\vec{\mathbf{p}})\tag{34}$$

In the ordinary differentiation of the product of two functions a(t) and b(t), we would have

$$\frac{\mathrm{d}}{\mathrm{d}t}(\mathrm{ab}) = \left(\frac{\mathrm{d}a}{\mathrm{d}t}\right)\mathrm{b} + \mathrm{a}\left(\frac{\mathrm{d}b}{\mathrm{d}t}\right) \tag{35}$$

The same rules apply if we differentiate a vector cross product. Thus

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\vec{r}\times\vec{p}\right) = \left(\frac{\mathrm{d}\vec{r}}{\mathrm{d}t}\right)\times\vec{p} + \vec{r}\times\left(\frac{\mathrm{d}\vec{p}}{\mathrm{d}t}\right) \tag{26}$$

Equation 36 can be simplified by noting that

$$\vec{v} = \frac{d\vec{r}}{dt}$$

so that

$$\left(\frac{d\vec{r}}{dt}\right) \times \vec{p} = \vec{v} \times \vec{p} = \vec{v} \times \left(m\vec{v}\right) = 0 \qquad (37)$$

This product is zero because the vectors  $\vec{v}$  and  $\vec{p} = m\vec{v}$  are parallel to each other, and the cross product of parallel vectors is zero. Thus Equation 36 becomes

$$\frac{\mathrm{d}}{\mathrm{d}t} \left( \vec{r} \times \vec{p} \right) = \vec{r} \times \frac{\mathrm{d}\vec{p}}{\mathrm{d}t} \tag{8}$$

With this result, let us return to Newton's law for linear momentum

$$\vec{F} = \frac{d\vec{p}}{dt} \tag{39}$$

As long as we do the same thing to both sides of an equation, it is still a correct equation. Taking the vector cross product  $\vec{r} \times$  on both sides gives

$$\vec{r} \times \vec{F} = \vec{r} \times \frac{d\vec{p}}{dt}$$
 (40)

Using Equation 38 in Equation 40 gives

$$\vec{\mathbf{r}} \times \vec{\mathbf{F}} = \frac{\mathbf{d}}{\mathbf{dt}} (\vec{\mathbf{r}} \times \vec{\mathbf{p}})$$
 (41)

Finally note that  $\vec{r}\times\vec{p}\,$  is the particle's angular momentum  $\vec{\ell}$  , thus

$$\vec{r} \times \vec{F} = \frac{d\vec{\ell}}{dt} \tag{42}$$

Equation (39) told us that the net linear force is equal to the time rate of change of linear momentum. Equation 42 tells us that something,  $\vec{r} \times \vec{F}$ , is equal to the time rate of change of angular momentum. What should we call this quantity  $\vec{r} \times \vec{F}$ ? The obvious name, from an angular analogy would be an *angular force*. Then we could say that the angular force is the time rate of change of angular momentum, just as the linear force is the time rate of change of linear momentum.

The world does not use the name angular force for  $\vec{r} \times \vec{F}$ . Instead it uses the name *torque*, and usually designates it by the Greek letter  $\tau$  ("tau")

$$torque \ \vec{\tau} \equiv \vec{r} \times \vec{F} \qquad definition \\ of torque \qquad (43)$$

With this naming, the angular analogy to Newton's second law is

$$\vec{\tau} = \frac{d\vec{\ell}}{dt} \begin{vmatrix} torque = rate \\ of change of \\ angular momentum \end{vmatrix}$$
(44)

## **ABOUT TORQUE**

To gain an intuitive picture of the concept of torque  $\vec{\tau} = \vec{r} \times \vec{F}$ , imagine that we have a bicycle wheel with a fixed axis, and push on the rim of the wheel with a force  $\vec{F}$  as shown in Figure (17). In (17a) the force  $\vec{F}$  is directed through the axis of the wheel, in this case the force has no lever arm  $r_{\perp}$ . In (17b), the force is applied above the axis, while in (17c) the force is applied below the axis.

Intuitively, you can see that the wheel will not start turning if you push right toward the axis. When you push above the axis as in (17b), the wheel will start to rotate counter clockwise. By our right hand convention this corresponds to an angular momentum directed up out of the paper.

In (17c), where we push below the axis, the wheel will start to rotate clockwise, giving it an angular momentum directed down into the paper.



Figure 17

Both a force  $\vec{F}$  and a lever arm  $r_{\perp}$  are needed to turn the bicycle wheel. The product  $r_{\perp}F$  is the magnitude of the torque  $\tau$  acting on the wheel.

#### Exercise 8

In Figure (17) we have separately drawn the vectors  $\vec{F}$  and  $\vec{r}$  for each diagram. Using the right hand rule for cross products, find the direction of  $\vec{\tau} = \vec{r} \times \vec{F}$  for each of these three diagrams.

If you did Exercise 8 correctly, you found that  $\vec{r} \times \vec{F} = 0$ for Figure (17a), that  $\vec{\tau} = \vec{r} \times \vec{F}$  pointed up out of the paper in (17b), and down into the paper in (17c). Thus we find that when we apply a zero torque as in (17a), we get zero change in angular momentum. In (17b) we applied an upward directed torque, and saw that the wheel would start to turn to produce an upward directed angular momentum. In (17c), the downward directed torque produces a downward directed angular momentum. These are all results we would expect from the equation  $\vec{\tau} = d\vec{\ell}/dt$ .

In our discussion of angular momentum, we saw that  $\vec{l} = \vec{r} \times \vec{p}$  had a magnitude  $|\vec{l}| = r_{\perp}p$  where  $r_{\perp}$  was the perpendicular lever arm. A similar result applies to torque. By the same mathematics we find that the magnitude of the torque  $\vec{\tau}$  produced by a force  $\vec{F}$  is

$$\left|\vec{\tau}\right| = r_{\perp}F \tag{45}$$

where  $r_{\perp}$  is the perpendicular lever arm seen in Figures (17b,c).

Intuitively, the best way to remember torque is to think of it as a force times a lever arm. To turn an object, you need both a force and a lever arm. In Figure (17a), we had a force but no lever arm. The line of action of the force went directly through the axis, with the result that the wheel did not start turning. In both cases (17b) and (17c), there was both a force and a lever arm  $r_{\perp}$ , and the wheel started turning.

To get the direction of the torque, to determine whether  $\vec{\tau}$  points up or down (and thus gives rise to an up or down angular momentum), use the right hand rule applied to the vector cross product  $\vec{\tau} = \vec{r} \times \vec{F}$ . A convention, which we will use in the next chapter on *Equilibrium*, is to say that a torque that points up out of the paper is a positive torque, while a torque pointing down into the paper is a negative one. With this convention, we see that the force in Figure (17b) is exerting a positive torque (and creating positive angular momentum), while the force in Figure (17c) is producing a negative torque (and creating negative angular momentum).

## CONSERVATION OF ANGULAR MOMENTUM

In our discussion of a system of particles in Chapter 11, we saw that if we had a system of many interacting particles, with internal forces  $\vec{F}_{i \text{ internal}}$  between the particles, as well as various external forces  $\vec{F}_{i \text{ external}}$ , we obtained the equation

$$\vec{F}_{external} = \frac{dP}{dt}$$
 (11-12)

where  $\vec{F}_{external}$  is the vector sum of all the external forces acting on the system, and  $\vec{P}$  is the vector sum of all the momenta  $\vec{p}_i$  of the individual particles. This result was obtained using Newton's third law and noting that all the internal forces cancel in pairs. In the case where there is no net external force acting on the system, then  $d\vec{P}/dt = 0$  and the total linear momentum of the system is conserved.

We can obtain a similar result for angular momentum by starting with the definition of the total angular momentum  $\vec{L}$  of a system as being the vector sum of the angular momentum of the individual particles  $\vec{\ell}_i$ 

$$\vec{L} \equiv \sum_{i} \vec{\ell}_{i}$$
definition of the  
total angular momenta  
of a system of particles
(46)

Differentiating Equation (46) with respect to time gives

$$\frac{d\vec{L}}{dt} = \sum_{i} \frac{d\vec{\ell}_{i}}{dt}$$
(47)

For an individual particle i, we have

$$\frac{d\vec{\ell}_{i}}{dt} = \vec{\tau}_{i} = \vec{r}_{i} \times \vec{F}_{i} \qquad \begin{array}{c} Equation \ 44\\ applied \ to\\ particle \ i \end{array} \tag{48}$$

where  $\vec{F}_i$  is the vector sum of the forces acting on the particle i. As shown in Figure (18), we can take  $\vec{r}_i$  to be the coordinate vector of the i-th particle. For this discussion, we can locate the origin of the coordinate system anywhere we want.

Substituting Equation (48) into Equation (47) gives

$$\frac{d\vec{L}}{dt} = \sum_{i} \frac{d\ell_{i}}{dt} = \sum_{i} \vec{r}_{i} \times \vec{F}_{i}$$

Now break the net force  $\vec{F}_i$  into the sum of the external forces  $\vec{F}_{i \text{ external}}$  and the sum of the internal forces  $\vec{F}_{i \text{ internal}}$ . This gives

$$\frac{d\vec{L}}{dt} = \sum_{i} \vec{r}_{i} \times \vec{F}_{i \text{ external}} + \sum_{i} \vec{r}_{i} \times \vec{F}_{i \text{ internal}}$$

$$= \sum_{i} \vec{\tau}_{i \text{ external}} + \sum_{i} \vec{\tau}_{i \text{ internal}}$$
(49)

Next assume that all the internal forces are equal and opposite as required by Newton's third law, *and* are directed toward or away from each other. In Figure (19) we consider a pair of such internal forces and note that both coordinate vectors  $\vec{r}_1$  and  $\vec{r}_2$  have the same perpendicular lever arm  $\vec{r}_{\perp}$ . Thus the equal and opposite forces  $\vec{F}_{1,2 \text{ external}}$  and  $\vec{F}_{2,1 \text{ internal}}$  create equal and opposite torques which cancel each other in Equation (49). The result is that all torques produced by internal forces cancel in pairs, and we are left with the general result

$$\vec{\tau}_{external} = \frac{d\vec{L}}{dt}$$
(50)

where  $\vec{\tau}_{external}$  is the vector sum of all the external torques acting on the system of particles, and  $\vec{L}$  is the vector sum of the angular momentum of all of the particles.



In order to define torque or angular momentum, we have to choose an axis or origin for the coordinate vectors  $\vec{r}_i$ . (Both torque and angular momentum involve the lever arm  $\vec{r}_{\perp}$  about that axis.) Equation 50 is remarkably general in that it applies no matter what origin or axis we choose. In general, choosing a different axis will give us different sums of torques and adifferent total angular momentum, but the new torques and angular momenta will still obey Equation 50.

In some cases, there is a special axis about which there is no external torque. In the bicycle wheel demonstrations where we stood on a rotating platform, the freely rotating platform did not contribute any external torques about it own axis, which we called the z axis. As long as we did not touch another person or some furniture, then the z component of the external torques were zero. Since Equation 50 is a vector equation, that implies

$$\tau_{z \text{ external}} = \frac{dL_z}{dt} = 0$$
(51)

and we predict that the z component of the total angular momentum (us and the bicycle wheel) should be unchanged, remain constant, no matter how we turned the bicycle wheel. This is just what we saw. Another consequence of Equation 50 is that if we have an isolated system of particles with no net external torque acting on it, then the total angular momentum will be unchanging, will be conserved. This is one statement of the law of conservation of angular momentum. Our derivation of this result relied on the assumption of Newton's third law that all internal forces are equal and opposite and directed toward each other. Since angular momentum is conserved on an atomic, nuclear and subnuclear scale of distance, where Newtonian mechanics no longer applies, our derivation is in some sense backwards. We should start with the law of conservation of angular momentum as a fundamental law, and show for large objects which obey Newtonian mechanics, the sum of the internal torques must cancel. This is the kind of argument we applied to the conservation of linear momentum in Chapter 11 (see Equation 11-14).



2Ī

(Note: when the wheel is being held up, we are looking at the under side.)



## **GYROSCOPES**

The gyroscope provides an excellent demonstration of the predictive power of the equation  $\vec{\tau} = d\vec{L}/dt$ . Gyroscopes behave in peculiar, non intuitive ways. The fact that a relatively straightforward application of the equation  $\vec{\tau} = d\vec{L}/dt$  predicts this bizarre behavior, provides a graphic demonstration of the applicability of Newton's laws from which the equation is derived.

## Start-up

For this discussion, a bicycle wheel with a weighted rim will serve as our example of a gyroscope. To weight the rim, remove the tire and wrap copper wire around the rim to replace the tire. The axle needs to be extended as shown in Figure (20).

As an introduction to the gyroscope problem, start with the bicycle wheel at rest, hold the axle fixed, and apply a force  $\vec{F}$  to the rim as shown in Figure (20). The force shown will cause the wheel to start spinning in a direction so that the angular momentum  $\vec{L}$  points to the right as shown. (Curl the fingers of your right hand in the direction of rotation and your thumb points in the direction of  $\vec{L}$ .)



Figure 20

Spinning up the bicycle wheel. Note that the resulting angular momentun  $\vec{L}$  points in the same direction as the applied torque  $\vec{\tau}$ .

The force  $\vec{F}$ , in Figure (20), produces a torque  $\vec{\tau} = \vec{r} \times \vec{F}$  that also points to the right as shown. (The right hand convention used here is to point your fingers in the direction of the first vector  $\vec{r}$ , curl them in the direction of the second vector  $\vec{F}$ , and your thumb points in the direction of the cross product  $\vec{r} \times \vec{F} = \vec{\tau}$ .)

When we start with the bicycle wheel at rest, and apply the right directed torque shown in Figure (20), we get a right directed angular momentum  $\vec{L}$ . Thus the torque  $\vec{\tau}$  and the resulting angular momentum  $\vec{L}$  point in the same direction. In addition, the longer we apply the torque, the faster the wheel spins, and the greater the angular momentum  $\vec{L}$ . Thus both the direction and magnitude of  $\vec{L}$  are consistent with the equation  $\vec{\tau} = d\vec{L}/dt$ .



Figure 25 Movie The gyroscope really works!

#### Precession

When we apply the equation  $\vec{\tau} = d\vec{L}/dt$  to a gyroscope that is already spinning, and apply the torque in a direction that is not parallel to  $\vec{L}$ , the results are not so obvious.

Suppose we get the bicycle wheel spinning rapidly so that it has a big angular momentum vector  $\vec{L}$ , and then suspend the bicycle wheel by a rope attached to the end of the axle as shown in Figure (21).

To predict the motion of the spinning wheel, the first step is to analyze all the external forces acting on it. There is the gravitational force mg which points straight down, and can be considered to be acting at the center of mass of the bicycle wheel, which is the center of the wheel as shown. Then there is the force of the rope which acts along the rope as shown. No other detectable external forces are acting on our system of the spinning wheel.

One thing we know about the force  $\vec{F}_{rope}$  is that it acts at the point labeled O where the rope is tied to the axle. If we take the sum of the torques acting on the bicycle wheel about the suspension point O, then  $\vec{F}_{rope}$  has no lever arm about this point and therefore contributes no torque. The only torque about the suspension point O is produced by the gravitational force mg whose lever arm is  $\vec{r}$ , the vector going from point O down the axle to the center of the bicycle wheel as shown in Figure (21).



#### Figure 21

Suspend the spinning bicycle wheel by a rope attached to the axle. The gravitational force  $m\tilde{g}$ has a lever arm  $\vec{r}$  about the axis O. This creates a torque  $\vec{\tau} = \vec{r} \times m\tilde{g}$  pointing into the paper.

## The formula for this gravitational torque $\vec{\tau}_{\rm g}$ is

$$\tau_{g} = \vec{r} \times m\vec{g} \qquad \begin{array}{l} \text{torque about point } O \\ \text{produced by the} \\ \text{gravitational force} \\ \text{on the bicycle wheel} \end{array} \tag{52}$$

The new feature of the gyroscope problem, which we have not encountered before, is that the torque  $\vec{\tau}$  does not point in the same direction as the angular momentum  $\vec{L}$  of the bicycle wheel. If we look at Figure (21), point the fingers of our right hand in the direction of the vector  $\vec{r}$ , and curl our fingers in the direction of the vector mg, then our thumb points down into the paper. This is the definition of the direction of the vector cross product  $\vec{r} \times mg$ . But the angular momentum  $\vec{L}$  of the bicycle wheel points along the axis of the wheel to the right in the plane of the paper. In order to view both the angular momentum vector  $\vec{L}$  and the torque vector  $\vec{\tau}$  in the same diagram, we can look down on the bicycle wheel from the celing as shown in Figure (22).

When we started the wheel spinning, back in Figure (20), the torque  $\vec{\tau}$  and angular momentum  $\vec{L}$  pointed in the same direction, and we had the simple result that the longer we applied the torque, the more angular momentum we got. Now, with the torque and angular momentum pointing in different directions as shown in Figure (22), we expect that the torque will cause a change in the direction of the angular momentum.



Figure 22

Looking down from the ceiling, the vector  $m\vec{g}$ points down into the paper and  $\vec{\tau} = \vec{r} \times m\vec{g}$ points to the top of the page. In this view we can see both the vectors  $\vec{L}$  and  $\vec{\tau}$ . To predict the change in  $\vec{L}$ , we start with the angular form of Newton's second law

$$\vec{\tau} = \frac{d\vec{L}}{dt}$$

and multiply through by the short (but finite) time interval dt to get

$$d\vec{L} = \vec{\tau} dt \tag{53}$$

Equation 53 gives us  $d\vec{L}$ , which is the change in the bicycle wheel's angular momentum as a result of applying the torque  $\vec{\tau}$  for a short time dt.

To see the effect of this change  $d\vec{L}$ , we will use some of the terminology we used in the computer prediction of motion. Let us call  $\vec{L}_{old}$  the old value of the angular momentum that the bicycle wheel had before the time interval dt, and  $\vec{L}_{new}$  the new value at the end of the time interval dt. Then  $\vec{L}_{new}$  will be related to  $\vec{L}_{old}$  by the equation

$$\vec{L}_{new} = \vec{L}_{old} + d\vec{L}$$
(54)

Using Equation 53 for  $d\vec{L}$  gives

$$\vec{L}_{new} = \vec{L}_{old} + \vec{\tau} dt$$
(55)

A graph of the vectors  $\vec{L}_{old}$ ,  $\vec{L}_{new}$ , and  $\vec{\tau}$  dt is shown in Figure (23). In this figure the perspective is looking down on the bicycle wheel, as in Figure (22).



**Figure 23** The vectors  $\vec{L}_{old}$ ,  $\vec{L}_{new}$  and  $\vec{\tau} dt$  as seen from the top view of Figure 22.

Since the torque  $\vec{\tau}$  is in the horizontal plane, the vector  $\vec{L}_{new} = \vec{L}_{old} + \vec{\tau} dt$  is also in the horizontal plane. And since  $\vec{\tau}$  and  $\vec{L}_{old}$  are perpendicular to each other,  $\vec{L}_{new}$  has essentially the same length as  $\vec{L}_{old}$ . What is happening is that the vector  $\vec{L}$  is starting to rotate counter clockwise (as seen from above) in the horizontal plane.

One final, important point. For this experiment we were careful to spin up the bicycle wheel so that before we suspended the wheel from the rope, the wheel had a big angular momentum pointing along its axis of rotation. When we apply a torque to change the direction of  $\vec{L}$ , the axis of the wheel and the angular momentum vector  $\vec{L}$  move together. As a result the axis of the bicycle wheel also starts to rotate counter clockwise in the horizontal plane. The bicycle wheel, instead of falling as expected, starts to rotate sideways.

Once the bicycle wheel has turned an angle  $d\theta$  sideways, the axis of rotation and the torque  $\vec{\tau}$  also rotate by an angle  $d\theta$ , so that the torque  $\vec{\tau}$  is still perpendicular to  $\vec{L}$  as shown in Figure (24). Since  $\vec{\tau}$  always remains perpendicular to  $\vec{L}$ , the vector  $\vec{\tau}$  dt cannot change the length of  $\vec{L}$ . Thus the angular momentum vector  $\vec{L}$  remains constant in magnitude and rotates or "precesses" in the horizontal plane. This is the famous precession of a gyroscope which is nicely demonstrated using the bicycle wheel apparatus of Figure (21).



**Figure 24** After each time step dt, the angular momentum vector  $\vec{L}$  (and the bicycle wheel axis) rotates by another angle  $d\theta$ .

To calculate the rate of precession we note from Figures (23) or (24) that the angle  $d\theta$  is given by

$$d\theta = \frac{\tau dt}{L} \tag{56}$$

where we use the fact that  $\tau dt$  is a very short length, and thus  $\sin(d\theta)$  and  $d\theta$  are equivalent. Dividing both sides of Equation 56 through by dt, we get

$$\frac{\mathrm{d}\theta}{\mathrm{d}t} = \frac{\tau}{\mathrm{L}} \tag{57}$$

But  $d\theta/dt$  is just the angular velocity of precession, measured in radians per second. Calling this precessional velocity  $\Omega_{\text{precession}}(\Omega \text{ is just a capital } \omega$ "omega"), we get

$$\Omega_{precession} = \frac{\tau}{L} \begin{bmatrix} precessional \\ angular velocity \\ of a gyroscope \end{bmatrix}$$
(58)

#### **Exercise** 9

A bicycle wheel of mass m, radius r, is spun up to an angular velocity  $\omega$ . It is then suspended on an axle of length h as shown in Figure (21). Calculate

- (a) the angular momentum L of the bicycle wheel.
- (b) the angular velocity of precession.

(c) the time it takes the wheel to precess around once (the period of precession). [You should be able to obtain the period of precession from the angular velocity of precession by dimensional analysis.]

(d) A bicycle wheel of total mass 1kg and radius 40cm, is spun up yo a frequency  $f = 2\pi\omega = 10$  cycles/sec. The handle is 30cm long. What is the period of precession in seconds? Does the result depend on rhe mass of the bicycle wheel?

If you try the bicycle wheel demonstration that we discussed, the results come out close to the prediction. Instead of falling as one might expect, the wheel precesses horizontally as predicted. There is a slight drop when you let go of the wheel, which can be compensated for by releasing the wheel at a slight upward angle.

If you look at the motion of the wheel carefully, or study the motion of other gyroscopes (particularly the air bearing gyroscope often used in physics lectures) you will observe that the axis of the wheel bobs up and down slightly as it goes around. This bobbing, or epicycle like motion, is called *nutation*. We did not predict this nutation because we made the approximation that the axis of the wheel exactly follows the angular momentum vector. This approximation is very good if the gyroscope is spinning rapidly but not very good if  $\vec{L}$  is small. Suppose, for example we release the wheel without spinning it. Then it simply falls. It starts to rotate, but along a different axis. As it starts to fall it gains angular momentum in the direction of  $\vec{\tau}$ . A more accurate analysis of the motion of the gyroscope can become fairly complex. But as long as the gyroscope is spinning fast enough so that the axis moves with  $\vec{L}$ , we get the simple and important results discussed above.

## **APPENDIX** Moment of Inertia and Rotational Kinetic Energy

In the main part of the text, we briefly discussed moment of inertia as the angular analogy to mass in the formula for angular momentum. As linear momentum  $\vec{p}$  of an object is its mass m times its linear velocity  $\vec{v}$ 

$$\vec{p} = m\vec{v}$$
 linear momentum (A1)

the angular momentum  $\vec{\ell}$  is the angular mass or moment of inertia I time the angular velocity  $\vec{\omega}$ 

$$\vec{\ell} = I\vec{\omega}$$
 angular momentum (A2)

In the simple case of a bicycle wheel, where all the mass is essentially out at a distance (r) from the axis of the wheel, the moment of inertia about the axis is

$$I = Mr^{2} \qquad \begin{array}{c} moment \ of \ inertia \\ of \ a \ bicycle \ wheel \end{array}$$
(A3)

where M is the mass of the wheel.

When the mass of an object is not all concentrated out at a single distance (r) from the axis, then we have to calculate the moment of inertia of individual parts of the object that are at different radii r, and tie together the various pieces to get the total moment of inertia. This usually involves an integration, like the one we did in Equations 21 through 23 to calculate the moment of inertia of a solid cylinder.

For topics to be discussed later in the text, the earlier discussion of moment of inertia is all we need. But there are topics, such as rotational kinetic energy and its connection to moment of inertia, which are both interesting, and can be easily tested in both lecture demonstrations and laboratory exercises. We will discuss these topics here.

## **ROTATIONAL KINETIC ENERGY**

Let us go back to our example, shown in Figure (3) repeated here, of a ball of mass m, on the end of a massless stick of length r, rotating with an angular velocity  $\omega$ . The speed v of the ball is given by Equation 11 as

and the ball's kinetic energy will be

 $v = r\omega$ 

kinetic  
energy = 
$$\frac{1}{2}mv^2 = \frac{1}{2}m(r\omega)^2$$
  
=  $\frac{1}{2}(mr^2)\omega^2$  (A4)

Since the ball's moment of inertia I about the axis of rotation is  $mr^2$ , we get as the formula for the ball's kinetic energy

kinetic  
energy = 
$$\frac{1}{2}I\omega^2$$
 analogous  
to  $1/2mv^2$  (A5)

We see the angular analogy working again. The ball has a kinetic energy, due to its rotation, which is analogous to  $1/2 mv^2$ , with the linear mass m replaced by the angular mass I and the linear velocity v replaced by the angular velocity  $\omega$ .



**Figure 3 repeated** *Mass rotating on the end of a massless stick.* 

If we have a bicycle wheel of mass M and radius r rotating at an angular velocity  $\omega$ , we can think of the wheel as being made up of a collection of masses on the ends of rods as shown in Figure (6) repeated here. For each individual mass m<sub>i</sub>, the kinetic energy is  $1/2 \text{ m}_i \text{v}^2$  where  $\text{v} = \text{r} \omega$  is the same for all the masses. Thus the total kinetic energy is

kinetic energy  
of bicycle wheel = 
$$\sum_{i} \frac{1}{2} m_{i} r^{2} \omega^{2}$$
  
=  $\frac{1}{2} r^{2} \omega^{2} \sum_{i} m_{i}$   
=  $\frac{1}{2} r^{2} \omega^{2} M$ 

where the sum of the masses  $\sum m_i$  is just the mass M of the wheel. The result can now be written

kinetic energy  
of bicycle wheel 
$$= \frac{1}{2} (Mr^2)\omega^2$$
  
 $= \frac{1}{2} I\omega^2$  (A6)

If we call Mr<sup>2</sup> the angular mass, or moment of inertia I of the bicycle wheel, we again get the formula  $1/2 I \omega^2$  for kinetic energy of the wheel. Thus we see that, in calculating this angular mass or moment of inertia, it does not make any difference whether the mass is concentrated at one point as in Figure (3), or spread out as in Figure (6). The only criterion is that the mass or masses all be out at the same distance r from the axis of rotation.



**Figure 6 repeated** Bicycle wheel as a collection of masses on the end of massless rods.

In most of our examples we will consider objects like bicycle wheels or hollow cylinders where the mass is essentially all at a distance r from the axis of rotation, and we can use the formula  $Mr^2$  for the moment of inertia. But often the mass is spread out over different radii and we have to calculate the angular mass. An example is a rotating shaft shown back in Figure (9), where the mass extends from the center where r = 0 out to the outside radius r = R.

Suppose we have an arbitrarily shaped object rotating an angular velocity  $\omega$  about some axis, as shown in Figure (A1). To find the moment of inertia, we will calculate the kinetic energy of rotation and equate that to  $1/2I\omega^2$  to obtain the formula for I. To do this we conceptually break the object into many small masses dm<sub>i</sub> located a distance r<sub>i</sub> from the axis of rotation as shown. Each dm<sub>i</sub> will have a speed v<sub>i</sub> = r<sub>i</sub>  $\omega$ , and thus a kinetic energy

kinetic energy  
of object 
$$= \sum_{i} \left( \frac{1}{2} m_{i} v_{i}^{2} \right)$$
$$= \sum_{i} \left( \frac{1}{2} m_{i} r_{i}^{2} \omega^{2} \right)$$
$$= \frac{1}{2} \omega^{2} \sum_{i} \left( m_{i} r_{i}^{2} \right)$$
$$= \frac{1}{2} \omega^{2} I \qquad (A7)$$

From Equation A7, we see that the general formula for moment of inertia is



Calculating the moment of inertia of an object about the axis of rotation.

In example 1, Equations 21 through 23, we showed you how to calculate the moment of inertia of a solid cylinder about its axis of symmetry. In that example we broke the cylinder up into a series of concentric shells of radius  $r_i$  and mass  $dm_i$ , calculated the moment of inertia of each shell  $(dm_i r_i^2)$ , and summed the results as required by Equation A7. As in most cases where we calculate a moment of inertia, the sum is turned into an integral.

In Exercise 3 which followed Example 1, we had you calculate the moment of inertia, about its axis of symmetry, of a hollow thick-walled cylinder. The calculation was essentially the same as the one we did in Example 1, except that you had to change the limits of integration. The following exercise gives you more practice calculating moments of inertia, and shows you what happens when you change the axis about which the moment of inertia is calculated.

#### Exercise A1

Consider a uniform rod of mass M and length L as shown in Figure (A2).

a) Calculate the moment of inertia of the rod about the center axis, labeled axis 1 in Figure (A2).

b) Calculate the moment of inertia of the rod about an axis that goes through the end of the rod, axis 2 in Figure (A2). About which axis is the moment of inertia greater? Explain why.



**Figure A2** *Calculating the moment of inertia of a long thin rod.* 

## COMBINED TRANSLATION AND ROTATION

In our discussion of the motion of a system of particles, we saw that the motion was much easier to understand if we focused our attention on the motion of the center of mass of the system. The simple feature of the motion of the center of mass, was that the effects of all internal forces cancelled. The center of mass moved as if it were a *point particle of mass M*, equal to the total mass of the system, subject to a force  $\vec{F}$  equal to the vector sum of all the *external forces* acting on the object.

When the system is a rigid object, we have a further simplification. The motion can then be described as the motion of the center of mass, plus rotation about the center of mass. To see that you can do this, imagine that you go to a coordinate system that moves with the object's center of mass. In that coordinate system, the object's center of mass point is at rest, and the only thing a rigid solid object can do is rotate about that point.

A key advantage of viewing the motion of a rigid object this way is that the kinetic energy of a moving, rotating, solid object is simply the kinetic energy of the center of mass motion plus the kinetic energy of rotation. Explicitly, if an object has a total mass M, and a moment of inertia  $I_{com}$  about the center of mass (parallel to the axis of rotation of the object) then the formula for the kinetic energy of the object is

kinetic energy  
of moving and 
$$= \frac{1}{2}MV_{com}^2 + \frac{1}{2}I_{com}\omega^2$$
 (A9)  
rotating object

where  $V_{com}$  is the velocity of the center of mass and  $\omega$  the angular velocity of rotation about the center of mass.

More important is the idea that motion can be separated into the motion of the center of mass plus rotation about the center of mass. To emphasize the usefulness of this concept, we will first consider an example that can easily be studied in the laboratory or at home, and then go through the proof of the equation.

## Example—Objects Rolling Down an Inclined Plane

Suppose we start with a cylindrical object at the top of an inclined plane as shown in Figure (A3), and measure the time the cylinder takes to roll down the plane. Since we do not have to worry about friction for a rolling object, we can use conservation of energy to analyze the motion.

If the cylinder rolls down so that its height decreases by h as shown, then the loss of gravitational potential energy is mgh. Equating this to the kinetic energy gained gives

mgh = 
$$\frac{1}{2}$$
 m v<sub>com</sub><sup>2</sup> +  $\frac{1}{2}$  I  $\omega^2$  (A10)

where m is the mass of the cylinder,  $v_{com}$  the speed of the axis of the cylinder, I the moment of inertia about the axis and  $\omega$  the angular velocity.

If the cylinder rolls without slipping, there is a simple relationship between  $v_{com}$  and  $\omega$ . We are picturing the rolling cylinder as having two kinds of motion—translation and rotation. The velocity of any part of the cylinder is the vector sum of  $\vec{v}_{com}$  plus the velocity due to rotation.

At the point where the cylinder touches the inclined plane, the rotational velocity has a magnitude  $v_{rot} = \omega r$ , and is directed back up the plane as shown in Figure (15). If the cylinder is rolling without slipping, the velocity of the cylinder at the point of contact must be zero, thus we have



**Figure A3** *Calculating the speed of an object rolling down a plane.* 

Thus we get for magnitudes

$$\omega r = v_{com}; \ \omega = v_{com}/r$$
 (A12)

Using Equation A12 in A10 gives

$$mgh = \frac{1}{2} mv_{com}^{2} + \frac{1}{2} I \frac{v_{com}^{2}}{r^{2}}$$
$$= \frac{1}{2} \left( m + \frac{I}{r^{2}} \right) v_{com}^{2}$$
(A13)

Let us take a look at what is happening physically as the cylinder rolls down the plane. In our earlier analysis of a block sliding without friction down the plane, all the gravitational potential energy mgh went into kinetic energy 1/2 mv<sup>2</sup><sub>com</sub>. Now for a rolling object, the gravitational potential has to be shared between the kinetic energy of translation  $1/2 \text{ mv}^2_{\text{com}}$  and the kinetic energy of rotation  $1/2 \text{ I}\omega^2$ . The greater the moment of inertia I, the more energy that goes into rotation, the less available for translation, and the slower the object rolls down the plane.

In our discussion of moments of inertia, we saw that for two cylinders of equal mass, the hollow thin-walled cylinder had twice the moment of inertia as the solid one. Thus if you roll a hollow and a solid cylinder down the plane, the solid cylinder will travel faster because less gravitational potential energy goes into the kinetic energy of rotation. You get to figure out how much faster in Exercise A2.





The velocity at the point of contact is the sum of the center of mass velocity and the rotational velocity. This sum must be zero if there is no slipping.

Before you work Exercise A2, think about this question. The technician who sets up our lecture demonstrations has a metal sphere, and does not know for sure whether the sphere is solid or hollow. (It could be a solid sphere made of a light metal, or a hollow sphere made from a more dense metal.) How could you find out if the sphere is solid or hollow?

#### **Exercise A2**

You roll various objects down the inclined plane shown in Figure (A3).

- (a) a thin walled hollow cylinder
- (b) a solid cylinder
- (c) a thin walled sphere
- (d) a solid sphere

and for comparison, you also slide a frictionless block down the plane:

(e) a frictionless block

For each of these, calculate the speed  $V_{com}$  after the object has descended a distance h. (It is easy to do all cases of this problem by writing the object's moment of inertia in the form  $I = \alpha MR^2$ , where  $\alpha = 1$  for the hollow cylinder, 1/2 for the solid cylinder, etc.) What value of  $\alpha$  should you use for the sliding block?

Writing your results in the form  $\vec{v}_{com} = \beta \sqrt{2gh}$  summarize your results in a table giving the value of  $\beta$  in each case. ( $\beta = 1$  for the sliding block, and is less than 1 for all other examples.)

#### Exercise A3 A Potential Lab Experiment

In Exercise A2 you calculated the speed  $v_{com}$  of various objects after they had descended a distance h. A block sliding without friction has a speed v given by mgh =  $1/2 \text{ mv}^2$ , or  $v = \sqrt{2gh}$ . The rolling objects were moving slower when they got to the bottom. For all heights, however, the speed of a rolling object is slower than the speed of the sliding block by the same constant factor. Thus the rolling objects moved down the plane with constant acceleration, but less acceleration than the sliding block. It is as if the acceleration due to gravity were reduced from the usual value g. Using this idea, and the results of Exercise A2, predict how long each of the rolling objects take to travel down the plane. This prediction can be tested with a stop watch.

## PROOF OF THE KINETIC ENERGY THEOREM

We are now ready to prove the kinetic energy theorem for rotational motion. If we have an object that is rotating while it moves through space, its total kinetic energy is the sum of the kinetic energy of the center of mass motion plus the kinetic energy of rotational motion about the center of mass. The proof is a bit formal, but shows what you can do by working with vector equations.

Consider a solid object, shown in Figure (A5), that is moving and rotating. Let  $\vec{R}_{com}$  be the coordinate vector of the center of mass of the object. We will think of the object as being composed of many small masses  $m_i$  which are located at  $\vec{R}_i$  in our coordinate system, and a displacement  $\vec{r}_i$  from the center of mass as shown. As we can see from Figure (A5), the vectors  $\vec{R}_{com}$ ,  $\vec{R}_i$  and  $\vec{r}_i$  are related by the vector equation

$$\vec{R}_i = \vec{R}_{com} + \vec{r}_i \tag{A14}$$

We can obtain an equation for the velocity of the small mass  $m_i$  by differentiating Equation A14 with respect to time

$$\frac{d\vec{R}_{i}}{dt} = \frac{d\vec{R}_{com}}{dt} + \frac{d\vec{r}_{i}}{dt}$$
(A15)



**Figure A5** Analyzing the motion of a small piece of an object.

which can be written in the form

$$\vec{V}_i = \vec{V}_{com} + \vec{v}_i \tag{A16}$$

where  $\vec{V}_i = d\vec{R}_i/dt$  is the velocity of  $m_i$  in our coordinate system,  $\vec{V}_{com} = d\vec{R}_{com}/dt$  is the velocity of the center of mass of the object, and  $\vec{v}_i = d\vec{r}_i/dt$  is the velocity of  $m_i$  in a coordinate system that is moving with the center of mass of the object.

The kinetic energy of the small mass  $m_i$  is

$$\frac{1}{2}m_{i}V_{i}^{2} = \frac{1}{2}m_{i}\vec{V}_{i}\cdot\vec{V}_{i}$$

$$= \frac{1}{2}m_{i}(\vec{V}_{com}+\vec{v}_{i})\cdot(\vec{V}_{com}+\vec{v}_{i})$$

$$= \frac{1}{2}m_{i}(V_{com}^{2}+2\vec{V}_{com}\cdot\vec{v}_{i}+v_{i}^{2})$$

$$= \frac{1}{2}m_{i}V_{com}^{2}+\frac{1}{2}m_{i}v_{i}^{2}+m_{i}\vec{V}_{com}\cdot\vec{v}_{i}$$
(A17)

The total kinetic energy of the object is the sum of the kinetic energy of all the small pieces  $m_i$ 

In two of the terms, we could take the common factor  $V_{com}$  outside the sum.



#### Figure A6

Here we moved the origin of the coordinate system to the center of mass.

Now the quantity  $m_i \vec{v}_i$  that appears in the last term of Equation A18 is the linear momentum of  $m_i$  as seen in a coordinate system where the center of mass is at rest. To evaluate the sum of these terms, let us choose a new coordinate system whose origin is at the center of mass of the object as shown in Figure (A6). In this coordinate system the formula for the center of mass of the small masses  $m_i$  is

$$\vec{r}_{com} = \sum_{i} m_{i} \vec{r}_{i} = 0$$
 (A19)

Differentiating Equation 19 with respect to time gives

$$\sum_{i} m_{i} \frac{d\vec{r}_{i}}{dt} = \sum_{i} m_{i} \vec{v}_{i} = 0$$
 (A20)

Equation A20 tells us that when we are moving along with the center of mass of a system of particles, the total linear momentum of the system, the sum of all the  $m_i \vec{v}_i$ , is zero.

Using Equation A20 in A18 gets rid of the last term. If we let  $M = \sum_{i} m_{i}$  be the total mass of the object, we get

$$\begin{array}{c} \text{total} \\ \text{kinetic} \\ \text{energy} \end{array} \right) = \frac{1}{2} M V_{\text{com}}^2 + \sum_{i} \frac{1}{2} m_i v_i^2 \quad (A2l)$$

Equation A21 applies to any system of particles, whether the particles make up a rigid object or not. The first term,  $1/2MV_{com}^2$  is the kinetic energy of center of mass motion, and  $\sum 1/2 m_i v_i^2$  is the kinetic energy as seen by someone moving along with the center of mass. If the object is solid, then in a coordinate system where the center of mass is at rest, the only thing the object can do is rotate about the center of mass. As a result the kinetic energy in that coordinate system is the kinetic energy of rotation. If the moment of inertia about the axis of rotation is  $I_{com}$ , then the total kinetic energy is  $1/2MV_{com}^2 + 1/2 I_{com}\omega^2$  where  $\omega$  is the angular velocity of rotation. This is the result we stated in Equation A9.
# Chapter 13 Equilibrium

When does a structure fall over, when does a bridge collapse, how do you lift a weight in a way that prevents serious injury to your back? We begin to answer such questions by applying Newton's laws to an object that has neither linear nor angular acceleration. The most interesting special case is when an object is at rest and will stay that way, when it is not about to tip over or collapse.

# EQUATIONS FOR EQUILIBRIUM

If the center of mass of an object is not accelerating, then we know that the vector sum of the external forces acting on it is zero. If the object has no angular acceleration, then the sum of the torques about any axis must be zero. These two conditions

sum of  
external 
$$\sum_{i} \vec{F}_{i \text{ external}} = 0$$
 (1)

sum of torques about any axis  $\sum_{i} \vec{\tau}_{i \text{ external}} = 0$ (2)

are what we will consider to be the required conditions for an object to be in *equilibrium*.

Equations 1 and 2 are a complete statement of the basic physics to be discussed in this chapter. Everything else will be examples to show you how to effectively apply these equations in order to understand and predict when an object will be in equilibrium. In particular we wish to show you some techniques that make it quite easy to apply these equations.



### Figure 1

Masses  $m_1$  and  $m_2$  suspended from a massless rod. At what position x do we suspend the rod in order for the rod to balance?

## **Example 1** Balancing Weights

As our first example, suppose we have a massless rod of length L and suspend two masses  $m_1$  and  $m_2$  from the ends of the rod as shown in Figure (1). The rod is then suspended from a string located a distance x from he left end of the rod. What is the distance x and how strong a force  $\vec{F}$  must be exerted by the string?

**Solution:** The first step is to sketch the situation and draw the forces involved, as we did in Figure (1). Our system will be the rod and the two masses. The external forces acting on this system are the two gravitational forces  $m_1 \vec{g}$  and  $m_2 \vec{g}$ , and the force of the string  $\vec{F}$ . Since all the forces are y directed, when we set the vector sum of these external forces to zero we have

$$\Sigma F_{v} = F - m_{1}g - m_{2}g = 0$$
 (3)

Thus we get for F

$$\mathbf{F} = (\mathbf{m}_1 + \mathbf{m}_2)\mathbf{g} \tag{4}$$

and we see that F must support the weight of the two masses.

To figure out where to suspend the rod, we use the condition that *the net torque produced by the three external forces must be zero*. Since a torque is a force times a lever arm about some axis, you have to choose an axis before you can calculate any torques. The important point in equilibrium problems is that *you can choose the axis you want*. We will see that by intelligently selecting an axis, we can simplify the problem to a great extent.

Our definition of a torque  $\vec{\tau}$  caused by a force  $\vec{F}$  is

$$\vec{t} = \vec{r} \times \vec{F} \tag{5}$$

where  $\vec{r}$  is a vector from the axis O to the point of application of the force  $\vec{F}$  as shown in Figure (2).

In this chapter we do not need the full vector formalism for torque that we used in the discussion of the gyroscope. Here we will use the simpler picture that the magnitude of a torque caused by a force  $\vec{F}$  is equal to the magnitude of F times the lever arm  $r_{\perp}$ , which is the distance of closest approach between the axis and the line of action of the force  $\vec{F}$  as shown in Figure (2). If the torque tends to cause a counter clockwise rotation, as it is in Figure (2), we will call this a positive torque. If it tends to cause a clockwise rotation, we will call that a negative torque. (In Figure (2), the vector  $\vec{r} \times \vec{F}$  points up out of the page. If the vector  $\vec{F}$  were directed to cause a clockwise rotation, then  $\vec{r} \times \vec{F}$  would point down into the paper. [It is good practice to check this for yourself.] Thus we are using the convention that torques pointing up are positive, and those pointing down are negative.)

Returning to our problem of the rod and weights shown in Figure (1), let us take as our axis for calculating torques, the point of suspension of the rod, labeled point O in Figure (1). With this choice, the force  $\vec{F}$ which passes though the point of suspension, has no lever arm about point O and therefore produces no torque about that point.

The gravitational force  $m_1 \vec{g}$  has a lever arm x about point O and is tending to rotate the rod counter clockwise. Thus  $m_1 \vec{g}$  produces a positive torque of magnitude  $m_1 gx$ . The other gravitational force  $m_2 \vec{g}$  has a lever arm (L - x) and is tending to rotate the rod clockwise. Thus  $m_2 \vec{g}$  is producing a negative torque magnitude  $-m_2 g(L - x)$ . Setting the sum of the torques about point O equal to zero gives

$$m_1 g x - m_2 g (L - x) = 0 \tag{6}$$

$$\mathbf{x} = \left(\frac{\mathbf{m}_2}{\mathbf{m}_1 + \mathbf{m}_2}\right) \mathbf{L} \tag{7}$$

Let us check to see that Equation 7 is a reasonable result. If  $m_1 = m_2$ , then we get x = L/2 which says that with equal weights, the rod balances in the center. If, in the extreme,  $m_1 = 0$ , then we get x = L, which tells us that we must suspend the rod directly over  $m_2$ , also a reasonable result. And if  $m_2 = 0$  we get x = 0 as expected.





We obtained Equation 7 by setting the torques about the balance point equal to zero. This choice had the advantage that the suspending force  $\vec{F}$  had no lever arm and therefore did not appear in our equations. We mentioned earlier that the condition for equilibrium was that the *sum of the torques be zero about any axis*. In Exercises 1 and 2 we have you select different axes about which to set the torques equal to zero. With these other choices, you will still get the same answer for x, namely Equation 7, but you will have two unknowns, F and x, and have to solve two simultaneous equations. You will see that we simplified the work by choosing the suspension point as the axis and thereby eliminating F from our equation.

### Exercise 1

If we choose the left end of the rod as our axis, as shown in Figure (3), then only the forces  $\vec{F}$  and  $m_2 \vec{g}$  produce a torque

(a) Is the torque produced by  $\vec{F}$  positive or negative?

(b) Is the torque produced by  $m_2 \vec{g}$  positive or negative?

(c) Write the equation setting the sum of the torques about the left end equal to zero. Then combine that equation with Equation 4 for F and solve for x. You should get Equation 7 as a result.

## **Exercise 2**

Obtain two equations for x and F in Figure (1) by first setting the torques about the left end to zero, then by setting the torques about the right end equal to zero. Then solve these two equations for x and see that you get the same result as Equation 7



# GRAVITATIONAL FORCE ACTING AT THE CENTER OF MASS

When we are analyzing the torques acting on an extended object, we can picture the gravitational force on the whole object as acting on the center of mass point. To prove this very convenient result, let us conceptually break up a large object of mass M into many small masses  $m_i$  as shown in Figure (4), and calculate the total gravitational torque about some arbitrary axis O. An individual particle  $m_i$  located a distance  $x_i$  down the x axis from our origin O produces a gravitational torque  $\tau_i$  given by

$$\tau_{i} = (m_{i}g)x_{i} \tag{8}$$

where  $m_i g$  is the gravitational force and  $x_i$  the lever arm. Adding up the individual torques  $\tau_i$  to obtain the total gravitational torque  $\tau_O$  gives

$$\tau_{O} = \sum_{i} \tau_{i} = \sum_{i} m_{i} g x_{i}$$

$$= g \sum_{i} m_{i} x_{i}$$
(9)



### Figure 4

Conceptually break the large object of mass M into many small pieces of mass  $m_i$ , located a distance  $x_i$  down the x axis from our arbitrary origin O.

But the sum  $\sum m_i x_i$  is by definition equal to M times the x coordinate of the center of mass of the object

$$MX_{com} \equiv \sum_{i} m_{i} x_{i}$$
 (10)

Using Equation 10 in 9 gives

$$\tau_{\rm O} = {\rm MgX}_{\rm com} \tag{11}$$

Equation 11 says that the gravitational torque about any axis O is equal to the total gravitational force Mg times the horizontal coordinate of the center of mass of the object. Thus the gravitational torque is just the same as if all of the mass of the object were concentrated at the center of mass point.

## **Exercise 3**

A wheel and a plank each have a mass M. The center of the wheel is attached to one end of a uniform beam of length L. A nail is driven through the center of mass of the plank and nailed into the other end of the beam as shown in Figure (5). Where do you attach a rope around the beam so that the beam will balance? Explain how you got your answer.



#### Figure 5

A wheel and a plank are attached to the ends of a uniform beam.

# TECHNIQUE OF SOLVING EQUILIBRIUM PROBLEMS

In our discussion of the balance problem shown in Figure (1), we saw that there were several ways to solve the problem. We always have the condition that for equilibrium the vector sum of the forces is zero  $(\Sigma_i \vec{F}_i = 0)$  and sum of the torques  $\vec{\tau}_i$  about any axis is zero  $(\Sigma_i \vec{\tau}_i = 0)$ . By choosing various axes we can easily get enough, or more than enough equations to solve the problem. If we are not careful about the way we do this however, we can end up with a lot of simultaneous equations that are messy to solve.

Our first solution of the equilibrium condition for Figure (1) suggests a technique for simplifying the solution of equilibrium problems. In Equation 6 we set to zero the sum of the torques about the balance point O shown in Figure (1) reproduced here. We wanted to calculate the position x of the balance point, and were not particularly interested in the magnitude of the force  $\vec{F}$ . By taking the torques about the point O where  $\vec{F}$  has no lever arm, F does not appear in our equation. As a result the only variable in Equation 6 is x, which can be immediately solved to give the result in Equation 7. As we saw in Exercises 1 and 2, if we chose the torques about any other point, both variables x and F appear in our equations.

We will now consider some examples and exercises that look hard to solve, but turn out to be easy if you take the torques about the correct point. The trick is to find a point that eliminates the unknown forces you do not want to know about.



**Example 3** Wheel and Curb

A boy is trying to push a wheel up over a curb by applying a horizontal force  $\vec{F}_{boy}$  as shown in Figure (6a). The wheel has a mass m, radius r, and the curb a height h as shown. How strong a force does the boy have to apply?

**Solution:** We will consider the wheel to be the object in equilibrium, and as a first step sketch all the forces acting on the wheel as shown in Figure (6b). We can treat this as an equilibrium problem by noting that as the wheel is just about to go up over the curb, there is no force between the bottom of the wheel and the road.

There is, however, the force of the curb on the wheel, labeled  $\vec{F}_{curb}$  in Figure (6b). We know the point at which  $\vec{F}_{curb}$  acts but we do not know off hand either the magnitude or direction of  $\vec{F}_{curb}$ , nor are we asked to find  $\vec{F}_{curb}$ .





A boy, exerting a horizontal force on the axle of a wheel, is trying to push the wheel up over a curb. How strong a force must the boy exert?



### Figure 6b

Forces on the wheel as the wheel is just about to go up over the curb.

13-5

**Figure 1 (Repeated)** *We can eliminate any force by a proper choice of axis.* 

We can eliminate the unknown force  $\vec{F}_{curb}$  by setting to zero the torques about the point O where the curb touches the wheel. Since  $\vec{F}_{curb}$  has no lever arm about this point, it will not appear in the resulting equation.

In Figure (6c) we have sketched the geometry of the problem. About the point O the force  $\vec{F}_{boy}$  has a lever arm (r-h) and is tending to cause a clockwise rotation about point O. Thus  $\vec{F}_{boy}$  is producing (by our convention) a negative torque, of magnitude  $F_{boy}$  (r-h). The gravitational force mg has a lever arm  $\ell$  shown in Figure (6c), and is tending to produce a counter clockwise rotation. Thus it is producing a positive torque of magnitude + mg $\ell$  about point O. Since there are no other torques about point O, setting the sum of the torques equal to zero gives

$$-F_{boy}(r-h) + mg\ell = 0$$

$$F_{boy} = mg\left(\frac{\ell}{r-h}\right)$$
(12)

We immediately see that if the curb is as high as the axle, if r = h, there is no finite force that will get the wheel over the curb.



**Figure 6c** *Geometry of the problem.* 



Figure 6d

The final slip in solving this problem is to relate the distance  $\ell$  to the wheel radius r and curb height h. As shown in Figure (6d), the right triangle from the axle to the curb has sides  $\ell$ , (r – h) and hypotenuse r. By the Pythagorean theorem we get

$$\mathbf{r}^2 = \ell^2 + (\mathbf{r} - \mathbf{h})^2$$

or

$$\ell = \sqrt{2 \operatorname{rh} - h^2} \tag{13}$$

which finishes the problem.

## **Exercise 4**

The direction of  $\vec{F}_{curb}$  is slightly off in Figure (6b). Explain what would happen to the wheel if  $\vec{F}_{curb}$  pointed as shown.



**Figure 7a** A frictionless rod is placed in a hemispherical frictionless bowl. What is the equilibrium position of the rod?



### Figure 7a

We simulated a frictionless rod in a hemispherical frictionless bowl by placing ball bearing rollers at one end of the rod and the edge of the "bowl". The rod always comes to rest at this angle.

# Example 4 Rod in a Frictionless Bowl

We include the problem here, first because it gives some practice with what we mean by a frictionless surface, but more importantly it is an example where we can gain considerable insight without solving any equations.

You place a frictionless rod of length 2r in a frictionless hemispherical bowl of radius r. Where does the rod come to rest? (Put in just enough friction to have it come to rest.) The situation is diagramed in Figure (7a). In Figure (7b), we have made a reasonably accurate simulation of the problem by using a semi circular piece of plastic for the bowl and placing small rollers on one end of the rod and one rim of the bowl to mimic the frictionless surfaces. In Figure (7c) we have sketched the forces acting on the rod. There is the downward force of gravity mg that acts at the center of mass of the rod, the force  $\vec{F}_b$  exerted by the bowl on the end of the rod, and the force  $\vec{F}_r$  exerted by the rim.

The idealization that we have a frictionless surface is equivalent to the statement that the surface can only exert normal forces, forces perpendicular to the surface. Thus the force  $\vec{F}_{h}$  exerted by the frictionless surface of the bowl is normal to the bowl and points toward the center of the circle defining the bowl. The force  $\vec{F}_r$  between the rim of the bowl and the frictionless rod must be perpendicular to the rod as shown.

Off hand we know nothing about the magnitude of the forces  $\vec{F}_b$  and  $\vec{F}_r$ , only their directions. If we extend the lines of action of  $\vec{F}_b$  and  $\vec{F}_r$  they will intersect at some point above the rod as shown in Figure (7c). If we set to zero the sum of the torques about this intersection point, where neither  $\vec{F}_b$  or  $\vec{F}_r$  has a lever arm, then neither  $\vec{F}_b$  or  $\vec{F}_r$  will contribute. The only remaining torque is that produced by the gravitational force  $m\vec{g}$ . If the rod is in equilibrium, then the torque produced by  $m\vec{g}$  about the intersection point must also be zero, with the result that the line of action of  $m\vec{g}$  must also pass through the intersection point as shown in Figure (7d). Thus the rod will come to rest when the center of mass of the rod lies directly below the intersection point of  $\vec{F}_b$  and  $\vec{F}_r$ . This result is nicely demonstrated by comparing the prediction, Figure (7d), with the experiment, Figure (7b).



## Figure 7c

Forces acting on the rod. Because the bowl is frictionless,  $\vec{F}_b$  is perpendicular to the surface of the bowl. Because the rod is frictionless  $\vec{F}_r$  is perpendicular to the rod.





For equilibrium, the center of mass must lie directly below the intersection point of  $\vec{F}_b$  and  $\vec{F}_r$ .

# **Exercise 5**

A spherical ball of mass m, radius r, is suspended by a string of length  $\ell$  attached to a frictionless wall as shown in Figure (8).

(a) show that the line of action of the tension force T (the line of the string) passes through the center of the ball.

(b) find the tension T.



**Figure 8** Ball suspended from a frictionless wall.

## **Exercise 6 Ladder Problem**

A ladder is leaning against a frictionless wall at an angle  $\theta$  as shown in Figure (9). Assume that the ladder is massless, and that a person of mass m is on the ladder.

The force between the ground and the bottom of the ladder can be decomposed into a normal component  $\vec{F}_n$ , and a horizontal component  $\vec{F}_f$  that can exist only if there is friction between the ladder and the ground.

It is traditional in introductory texts to say that the ladder will start to slip if the friction force  $\vec{F}_f$  exceeds a value of  $\mu F_n$  where  $\mu$  is called the "coefficient of static friction". This idea is reasonable in that as the normal force  $\vec{F}_n$  increases, so does the gripping or friction force  $\vec{F}_f$ . However the coefficient  $\mu$  depends so much upon the circumstances of the particular situation, that the theory is not particularly useful. What, for example, should you use for the value of  $\mu$  if the ends of the ladder sink down into the ground?

However for the sake of this problem, assume that the ladder will just start to slip when  $F_f = \mu F_n$ . Assume that  $\mu$  has the value  $\mu = 1/\sqrt{3} = .557$ .

(a) at what angle  $\theta$  would you place the ladder so that it will not start to slip until the person climbing it just reaches the top?

(b) at what angle  $\theta$  would you place the ladder so that it will not start to slip until the person has gone half way up?



**Figure 9** Ladder leaning against a frictionless wall.

## Example 5 A Bridge Problem

A bridge is constructed from massless rigid beams of length  $\ell$ . The ends of the beams are connected by a single large bolt that acts more or less like a big hinge. As a result the only forces you can have in each beam is either tension or compression (i.e. each beam either pulls or pushes along the length of the beam.) The idea is to be able to calculate the tension or compression force in any of the beams when a load is placed on the bridge.

In this example, we will place a mass m in the center of the right most span as shown in Figure (10a). To illustrate the process of calculating tension or compression in the beams, we will calculate the force in the upper left hand beam. For now we will assume that the beam is under tension and exerts a force  $\vec{T}_d$  on joint d as shown. If it turns out that the beam is under compression, then the magnitude of  $T_d$  will turn out to be negative. Thus we do not have to know ahead of time whether the beam is under tension or compression.

**Solution:** When you have a statics problem involving an object with a lot of pieces, and you want to calculate the force in one of the pieces, the first step is to isolate part of the object and consider that as a separate system with external forces acting on it. In Figure (10b) we have chosen as our isolated system the part of the bridge made from the girders that have been drawn in heavy lines. The external forces acting on this isolated system are the gravitational force  $m\vec{g}$  acting on the mass m, the supporting force  $\vec{F}_2$  that holds up the right end of the bridge, (we will assume that the ends of the bridge are free to slide back and forth, so that the supporting forces  $\vec{F}_1$  and  $\vec{F}_2$  point straight up). In addition the tension (or compression) forces we have labeled  $\vec{T}_d$ ,  $\vec{T}_{c1}$  and  $\vec{T}_{c2}$  are also acting on our isolated section of the bridge.

Looking at Figure (10b), it is immediately clear that the forces we do not want to know anything about are the tension forces  $\vec{T}_{c1}$  and  $\vec{T}_{c2}$ . We can eliminate these forces by setting to zero the torques about the joint labeled c. Using our convention that counter clockwise torques are positive and clockwise ones are negative, we get

$$\Sigma T_{aboutc} = F_2(2\ell) - mg\left(\frac{3}{2}\ell\right) + T_d\left(\frac{\sqrt{3}}{2}\ell\right) = 0$$
(13)

where we used the fact that  $T_d$ 's lever arm h is the altitude of an equilateral triangle.

In Equation 13 we have two unknowns,  $F_2$  and  $T_d$  Thus we need another equation. If we take the bridge as a whole, and calculate the torques on about the point a (to eliminate the force  $F_1$ ), we get

$$\Sigma_{\text{T}_{\text{about a}}} = F_2(3\ell) - mg\left(\frac{5}{2}\ell\right) = 0 \quad (14)$$

Solving Equation 14 for  $F_2$  gives

$$F_2 = \frac{5}{6} mg$$
 (15)

Using this value in Equation 13 gives

$$T_{d} = \frac{-mg}{3\sqrt{3}}$$
(18)

The minus (–) sign indicates that the beam is under compression.

## **Exercise 7**

Find the tension (or compression in the beam that goes from point (d) to point (e) in the bridge problem of Figure (10a).



**Figure 10a** *Bridge with a truck on the last span.* 



**Figure 10b** *Finding the tension in the span from b to d.* 

## **Exercise 8 Working with Rope**

As most sailors know, if you use rope correctly, you can create very large tension forces without exerting that strong a force yourself. Suppose, for example, you wish to make a raft out of two long logs with two short spacer planks between them as shown in Figure (11a). You wish to hold the raft together with a rope around the center as shown.

The first step is to tie, as tightly as you can, the logs together as shown in the end view of Figure (11b). Then take another piece of rope and tie it as tightly as you can as shown in Figure (11c). If you do a reasonably good job, you can create a large tension in the rope holding the logs together.

To analyze the problem, let  $T_1$  be the tension in the rope holding the logs together, and  $T_2$  the tension in the line between the ropes as shown in Figure (11d). For this problem, assume that angle  $\theta$  in Figure (11d) is 5 degrees, and the tension  $T_2$  that you could supply in winding the line around the ropes was 200 newtons (enough force to lift a 20 kilogram mass). What is the tension  $T_1$  in the ropes holding the logs together?



## Figure 11a

Constructing a raft by tying two logs together, with wood spacers.



**Figure 11b** End view of raft.



**Figure 11c** *Tightening the rope.* 



**Figure 11d** *Tensions in the ropes.* 

# LIFTING WEIGHTS AND MUSCLE INJURIES

The previous exercise on tying a raft together illustrates the fact that with some leverage, you can create large tensions in a rope. Similar large forces can exist in your muscles when you lift weights, particularly if you do not lift the weights properly.

To illustrate the importance of lifting heavy objects correctly, consider the sketch of Figure (12) showing a shopper holding a funny looking 10 kg shopping bag out at arms length. We wish to determine the forces that must be exerted on the backbone and by the back muscles in order to support this extra weight.

To analyze the forces, think of the upper body and arm as essentially a rigid object supported by the backbone and back muscles as shown. Since we are interested in the extra forces required to lift the weight, we will ignore the weight of the upper body itself.

The external forces acting on the upper body are the upward compressional force  $\vec{F}_b$  acting on the backbone, and the downward force  $\vec{F}_m$  acting at the point where the back muscle is attached to the thighbone. (This is in reaction to the upward force exerted on the thigh bone by the contacting back muscle.) There is also the weight  $m\vec{g}$  of the shopping bag. We are letting L be the horizontal distance from the backbone to the shopping bag, and  $\ell$  the separation between the point where the thigh bone pushes up on the backbone and pulls down on the back muscle.

If we want to solve for the force  $\vec{F}_m$  exerted by the back muscle we can eliminate  $\vec{F}_b$  from our equation by setting to zero the sum of the torques about point b, the point where the thigh bone contacts the backbone. We get

$$\Sigma \tau_{about b} = F_m \ell - MgL = 0$$

$$F_m = mg\left(\frac{L}{\ell}\right)$$
(9)

We see that the back muscle has to pull down with a force that is a factor of  $L/\ell$  times greater than the weight mg of the shopping bag. With  $\ell = 2 \text{ cm}$ , you see that if you hold the shopping bag out at arms length, say L = 80 cm, then  $F_m$  is  $L/\ell = 40$  times as great as the weight of the shopping bag. For you to lift a 10 kg bag at arms length, your back is essentially lifting a 10 kg  $\times 40 = 400$  kg mass, which has a weight of almost half a ton! If instead you pulled your arm in so that L was only 20 cm, then  $F_m$  drops 1/4 of its original value. Do your back a favor and do not lift heavy objects out at arms length.

# Exercise 9

Write a single equation that allows you to solve for the compressional force  $\vec{F}_b$  exerted on the backbone, as shown in Figure (12).



# **Exercise 10**

Figure (13) shows the structure of the lower leg and foot. Assume that this person is raising her heel a bit off the ground, so that her foot is touching the floor at only one place, namely the point labeled P in the diagram. Also assume that the calf muscle is attached to the foot bones at the point labeled (a), and that the leg bone acts at the pivot point (b). If her mass is 60 kilograms, what must be the forces exerted by the calf muscle and the leg bone? Compare the strength of these forces with her weight. (This and the next problem adapted from Halliday & Resnick.)

# Exercise 11

The arm in Figure (14) is holding a 20 kilogram mass. The arm pivots around the points marked with a small black circle. What is the compressional force on the bones in this joint? (Neglect the mass of the arm.)









# Chapter 14 Oscillations and Resonance

Oscillations and vibrations play a more significant role in our lives than we realize. When you strike a bell, the metal vibrates, creating a sound wave. All musical instruments are based on some method to force the air around the instrument to oscillate. Oscillations from the swing of a pendulum in a grandfather's clock to the vibrations of a quartz crystal are used as timing devices. When you heat a substance, some of the energy you supply goes into oscillations of the atoms. Most forms of wave motion involve the oscillatory motion of the substance through which the wave is moving.

Despite the enormous variety of systems that oscillate, they have many features in common, features exhibited by the simple system of a mass on a spring. As a result we will focus our attention on the analysis of the motion of a mass on a spring, describing ways in which other forms of oscillation are similar.

# **OSCILLATORY MOTION**

Suspend a mass on the end of a spring as shown in Figure (1), gently pull the mass down and let go. The mass will oscillate up and down about the equilibrium position. How do we describe the kind of motion we are looking at?

One of the best ways to see what kind of motion we are dealing with is to perform the demonstration illustrated in Figure (2a). In that demonstration, we place a rotating wheel beside an oscillating mass, and view the two objects via a TV camera set off to the side as shown. A white tape is placed around the mass, and a short stick is mounted on the rotating wheel as seen in the edge view, Figure (2b). This edge view is the one displayed by the TV camera.

The wheel is mounted on a variable speed motor, which allows us to adjust the angular velocity of the  $\omega$  of the wheel so that the wheel goes around once in precisely the same length of time it takes the mass to bob up and down once. The height of the wheel is adjusted so that when the mass is at rest at its equilibrium position, the white stripe on the mass lines up with the axis of the wheel. As a result if the stick is in a horizontal position (3 o'clock or 9 o'clock) and the mass is at rest, the stick and the white stripe line up on the television image.



# Figure 1

Mass suspended from a spring. If you gently pull the mass down and release it, it will oscillate up and down about the equilibrium position. Now pull the mass down so that the white stripe lines up with the lowest position of the stick the same height as the stick when the stick is at the bottom position (6 o'clock). Start the motor rotating at the correct frequency and release the mass when the stick is at the bottom. If you do this just right, some practice may be required, you will see in the television picture that the white stripe and the stick move up and down together as if they were a single object.

From this demonstration we conclude that *the up and down oscillatory motion of the mass is the same as circular motion viewed sideways*. As a result we can use what we know about circular motion to understand oscillatory motion. As a start, we will say that the oscillatory motion has an angular frequency  $\omega$  that is the same as the angular velocity  $\omega$  of the rotating wheel when the mass and the stick go up and down together.



## Figure 2a

Lecture setup for comparing the oscillation of a mass on a spring with a rotating wheel. A stick is mounted on the rotating wheel, and a TV camera off to the side provides a side view of the oscillating mass and rotating wheel.



## Figure 2b

Side view of the oscillating mass and rotating wheel, as seen by the TV camera. When the motor is adjusted to the correct frequency, the mass and the stick are observed to move up and down together.

# THE SINE WAVE

There is another way to picture the sideways view of circular motion. As more or less a thought experiment, suppose that we take our rotating wheel with a stick shown in Figure (2a) and shine a parallel beam of light at it, sideways, as shown in Figure (3). Now picture a truck with a big billboard mounted on the back, driving away from the light source as shown on the right side of Figure (3). The image of the stick will move up and down on the billboard as the truck moves forward.

Finally picture the line traced out in space by the image of the stick on the moving billboard. The image is going up and down with a frequency  $\omega$  and moves forward at a speed v, the speed of the truck. The result is an undulating curve we call a *sine wave*.

To make this definition of the sine wave more specific, assume the truck crosses the point x = 0 just when the angle  $\theta$  of the stick is zero as shown in Figure (3). Let us imagine that the truck drives at a speed  $v = \omega$ , so that

```
the distance x = vt = \omega t that the truck has travelled is
the same as the angular distance \theta = \omega t that the stick
has travelled. Finally let the radius of the circle around
which the stick is travelling be r = 1, so that the
undulating curve goes up to a maximum value y = +1
and down to a minimum value y = -1. With these
conditions, the curve traced out is the mathematical
function
```

$$y = \sin \theta = \sin (\omega t)$$
 (1)

Let us remove the truck and billboard and look at the sine curve itself more carefully as shown in Figure (4). The horizontal axis of the sine curve is the angular distance  $\theta = \omega t$  that the rotating stick has travelled. Starting at 0 when  $\theta = 0$ , the sine curve completes one full cycle or undulation just when the wheel has gone around once and  $\theta = 2\pi$ . Thus one cycle of a sine wave goes from 0 to  $2\pi$  as shown in Figure (4). The sine wave reaches a maximum height at  $\theta = \pi/2$ , goes back to zero again at  $\theta = \pi$ , has a minimum value at  $3\pi/2$  and completes the cycle at  $\theta = \omega t = 2\pi$ .



Project the image of the stick onto the back of a truck moving at a speed  $v = \omega t$ , and the image traces out a sine wave.



**Figure 4** The sine curve  $y = \sin \theta = \sin \omega t$ .

## **Exercise 1**

Somewhere back in the dim past, you learned that  $\sin \theta$  was the ratio of the opposite side to the hypotenuse in a right triangle. Applied to Figure (5), this is

$$\sin \theta \equiv y/r \tag{2}$$

Show that this older definition of  $\sin \theta$ , at least for angles  $\theta$  between 0 and  $\pi/2$ , is the same as the definition of  $\sin \theta$  we are using in Figure (3) and (4).

As you can see from Exercise 1, our rotating stick picture of the sine wave is mathematically equivalent to the definition of  $\sin \theta$  you learned in your first trigonometry class. What may be new conceptually is the dynamic aspect of the definition. Figures (3) and (4) connect rotational motion to oscillatory motion and to the shape of a sine wave. The relationship between the static picture and the dynamic one is that the angular distance  $\theta$  is equal to the angular velocity  $\omega$  times the elapsed time t.





**Figure 5** *Our old definition of sin*  $\theta$ .

The basic question for the dynamic picture is how long does one oscillation take. The time for one oscillation is called the period T of the oscillation. We can therefore ask what is the period T of an oscillation whose angular velocity, or angular frequency is  $\omega$ .

The solution to this problem is to note that the sine wave completes one cycle when  $\theta = 2\pi$ . But  $\theta$  is just the angular distance  $\omega t$ . Thus, if t = one period T, we have

$$\theta = \omega T = 2\pi$$

$$T = \frac{2\pi}{\omega}$$
period of
a sinusoidal
oscillation
(3)

To remember formulas like Equation 3, we can use the same set of dimensions we used in our discussion of angular motion. If we remember the dimensions

$\omega \frac{radians}{second}$	angular frequency
$2\pi \frac{radians}{cycle}$	
$T \frac{seconds}{cycle}$	period
$f \frac{cycles}{second}$	frequency

then we can go back and forth between the quantities  $\omega$ , T and f simply by making the dimensions come out right.

For example

$$T \frac{\sec}{\text{cycle}} = \frac{2\pi \frac{\text{radians}}{\text{cycle}}}{\omega \frac{\text{radians}}{\text{sec}}} = \frac{2\pi}{\omega} \frac{\sec}{\text{cycle}}$$
$$f \frac{\text{cycles}}{\sec} = \frac{\omega \frac{\text{radians}}{\sec}}{2\pi \frac{\text{radians}}{\text{cycle}}} = \frac{\omega}{2\pi} \frac{\text{cycles}}{\sec}$$
$$T \frac{\sec}{\text{cycle}} = \frac{1}{c \text{ cycles}} = \frac{1}{f} \frac{\sec}{\text{cycle}}$$

sec

We repeated these dimensional exercises, because it is essential that you be able to easily go back and forth between quantities like frequency, angular frequency, and period.

## Exercise 2

(a) A spring is vibrating at a rate of 2 seconds per cycle. What is the angular velocity  $\omega$  of this oscillation?

(b) What is the period of oscillation, in seconds, of an oscillation where  $\omega = 1$ ?



**Figure 6** 

Mass with springs on an air cart. We take the equilibrium position to be x = 0.

### **Exercise 3**

As shown in Figure (6), an air cart sitting on an air track has springs attached to the ends of the track as shown. We are taking x = 0 to be the equilibrium position of the cart. It turns out that the car oscillates back and forth with the same kind of sinusoidal motion as the mass on the end of a spring shown in Figures (1) and (2).

Assume that the mathematical formula for the coordinate x of the cart is

$$x = x_0 \sin \omega t$$

where

$$x_0 = 4 \text{ cm}; \quad \omega = 3 \frac{\text{radians}}{\text{sec}}$$

(a) Figure (7) is an x–t graph of the position of the air cart in Figure (6). We have drawn in the sine curve, and drawn tick marks at important points along the x and t axis. On the x axis, the tick marks are at the maximum and minimum values of x. On the t axis, the tick marks are at 1/4, 1/2, 3/4 and 1 complete cycle. Insert on the graph, the numerical values that should be associated with these tick marks.

(b) Where will the cart be located at the time  $t = 300 \pi$  seconds?



**Figure 7** The x-t graph for the motion of the air cart in Figure 6.

# **Phase of an Oscillation**

Our sine waves, by definition, begin at 0 when  $\theta = 0$ . This is equivalent to saying that in Figure (2), the truck crossed x = 0 at the same instant that the rotating wheel crossed  $\theta = 0$ . We did not have to make this choice, the rotating wheel could have been at any angle  $\phi$  when the truck crossed t = 0 as shown in Figure (8). If the stick were up at an angle  $\phi$  when the truck crossed the zero of the horizontal axis, we say that the resulting sine wave has a phase  $\phi$ . The formula for the resulting sine curve is

$$y = \sin(\omega t + \phi) \qquad \begin{array}{l} \phi \text{ is the} \\ phase angle \end{array}$$
(4)

You can see from this equation that at t = 0, the angle of the sine wave is  $\theta = \omega t + \phi = \phi$ 



The phase angle  $\phi$ .

In Figure (9) we have sketched the sine wave for several different phases. At a phase  $\phi = \pi/2$ , the wave starts at 1 for  $\theta = 0$  and goes down to 0 at  $\theta = \pi/2 = 90^{\circ}$ . This is just what  $\cos \theta$  does, and we have what is called a cosine wave. We have

$$\cos\left(\omega t\right) = \sin\left(\omega t + \pi/2\right) \tag{6}$$

When the phase angle gets up to  $\pi$  or 180°, the sine wave is reversed into  $a - \sin(\omega t)$  wave. At  $\phi = 2\pi$  we are back to the sine wave we started with.



#### Figure 9

Various phases of the sine wave. When  $\phi = \pi/2$ , the wave is called a cosine wave. (It matches the definition of a cosine, which starts out at 1 for  $\theta = 0$ . At a phase angle  $\phi = \pi$  or  $180^{\circ}$ , the sine wave has reversed and become  $-\sin(\omega t)$ . At  $\phi = 2\pi$ , we are back to a sine wave again.

## **Exercise 4**

In trigonometry class, or somewhere perhaps, you were given the trigonometric identity

sin(a+b) = sin a cos b + sin b cos a

Use this result to show that

sin	$(\omega t + \pi/2)$	) =	cosωt	(7a	i)
-----	----------------------	-----	-------	-----	----

$\sin(\omega t + \pi) = -\sin(\omega t)$ (7)	n(oot + :	sin (wt	c) =	= -sin ωt	(	7	b
--	-----------	---------	------	-----------	---	---	---

$$\sin(\omega t + 3\pi/2) = -\cos \omega t \tag{7c}$$

$$\sin(\omega t + 2\pi) = \sin \omega t \tag{7d}$$

These are the results graphed in Figure (9).



If the cart is displaced a distance x from equilibrium, there is a restoring force F = -kx.



Restoring force for a mass on a spring.

# MASS ON A SPRING; ANALYTIC SOLUTION

Let us now apply Newton's laws to the motion of a mass on a spring and see how well the results compare with the sinusoidal motion we observed in the demonstration of Figure (2). In our analysis of the spring pendulum in Chapter 9 (see Figure 9-4) we saw that the spring exerted a force whose strength was linearly proportional to the amount the spring was stretched, a result known as Hooke's law.

When we have the one dimensional motion of a mass oscillating about its equilibrium position, as illustrated in Figure (10), then we get a very simple formula for the net force on the object. If we displace the cart of Figure (10) by a distance x from equilibrium, there is a restoring force whose magnitude is proportional to x pushing the cart back toward the equilibrium position.

We can completely describe this restoring force by the formula

$$F = -kx$$
Hooke's law
restoring force
(8)

If x = 0, the cart is at its equilibrium position and there is no force. If x is positive, as in Figure (10), the restoring force is negative, pointing back to the equilibrium position. And if x is negative, the restoring force points in the positive direction. All these cases are handled by the formula F = -kx.

For a mass bobbing up and down on a spring, shown in Figure (11), there is both a gravitational and a spring force acting on the mass. But if you measure the net force starting from the equilibrium position, you still get a linear restoring force. The net force is always directed back toward the equilibrium position and has a strength proportional to the distance x the mass is away from equilibrium. Thus Equation 8 still describes the net force on the mass.

## **Exercise 5**

Describe experiments you could carry out in the laboratory to measure the force constant k for the air cart setup of Figure (10). To do this you are given the air cart setup, a string, a pulley, and some small weights.

Sketch the setup you would use to make the measurements, and include some simulated data to show how you would obtain a numerical value of k from your data. (This is the kind of exercise you would do ahead of time if you planned to do a project studying the oscillatory motion of the air cart and spring system.

To apply Newton's laws to the problem of the oscillating mass, let x(t) be the displacement from equilibrium of either the air cart of Figure (10) or the mass of Figure (11). The velocity  $v_x$  and the acceleration  $a_x$  of either the cart or mass is

$$v_{x} \equiv \frac{dx(t)}{dt}$$
(9)

$$a_{x} \equiv \frac{dv_{x}}{dt} = \frac{d^{2}x(t)}{dt^{2}}$$
(10)

The x component of Newton's second law becomes

$$F_{x} = ma_{x}$$
$$-kx = m\frac{d^{2}x}{dt^{2}}$$
(11a)

where we used Hooke's law, Equation 8 for  $F_x$ .

The result, Equation 11a, involves both the variable x(t) and its second derivative  $d^2x/dt^2$ . An equation involving derivatives is called a *differential equation*, and one like Equation 11a, where the highest derivative is the second derivative, is called a *second order differential equation*. Differential equations are harder to solve than algebraic equations like  $x^2 = 4$ , because the answer is a function or a curve, rather than simple numbers like  $\pm 2$ .

When working with differential equations, there is a traditional form in which to write the equation. The highest derivative is written to the far left, all terms with the unknown variable are put on the left side of the equation, and the coefficient of the highest derivative is set to one. (A reason for this tradition is that only a few differential equations have been solved. If you write them all in a standard form, you may recognize the one you are working with.) Putting Equation 11a in the standard form by dividing through by m and rearranging terms gives

$$\frac{d^2x}{dt^2} + \frac{k}{m}x = 0$$
(11)

A standard way to solve a differential equation is to guess the answer, and then plug your guess into the equation to see if the guess works. A course in differential equations basically teaches you how to make good guesses. In the absence of such a course, we have to use whatever knowledge we have about the system in order to make as good a guess as we can. That is why we did the demonstration of Figure (2). In that demonstration we saw the oscillating mass moved the same way as a stick on a rotating wheel, when the wheel is viewed sideways. We then saw that this sideways view of rotating motion is described by the mathematical function  $sin(\omega t)$ . From this we suspect that a good guess for the function x(t) may be

$$\mathbf{x}(\mathbf{t}) = \sin\left(\boldsymbol{\omega}\mathbf{t}\right) \quad initial \; guess \tag{12}$$

In order to see if this guess is any good, we need to substitute values of x and  $d^2x/dt^2$  into Equation 11. To do this, we need derivatives of  $x = \sin(\omega t)$ .

From your calculus course you learned that

$$\frac{\mathrm{d}}{\mathrm{d}t}\sin\left(\omega t\right) = \omega\cos\left(\omega t\right) \tag{13}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}\cos\left(\omega t\right) = -\omega\sin\left(\omega t\right) \tag{14}$$

Thus if we start with

$$\mathbf{x} = \sin\left(\boldsymbol{\omega}t\right) \tag{12}$$

we have

$$\frac{dx}{dt} = \omega \cos(\omega t)$$

$$\frac{d^2x}{dt^2} = \frac{d}{dt} \left( \frac{dx}{dt} \right) = \frac{d}{dt} \left( \omega \cos \omega t \right)$$

$$= \omega \left( -\omega \sin \omega t \right)$$
(15)

$$\frac{d^2x}{dt^2} = -\omega^2 \sin\left(\omega t\right)$$
(16)

To check our guess  $x = \sin(\omega t)$ , we substitute the values of x from Equation 12 and  $d^2x/dt^2$  from Equation 16 into the differential Equation 11. We get

$$\frac{d^2 x}{dt^2} + \frac{k}{m} x = 0$$

$$\left[ -\omega^2 \sin\left(\omega t\right) \right] + \frac{k}{m} \left[ \sin\left(\omega t\right) \right] = 0 \qquad (17)$$

We put the question mark over the equal sign, because the question we want to answer is whether  $x = \sin \omega t$ can be a solution to our differential equation. Can the sum of these two terms be made equal to zero as required by Equation 11?

The first thing we note in Equation 17 is that the function  $(\sin \omega t)$  cancels. This is encouraging, because if we ended up with two different functions of time, say a sin  $\omega t$  in one term and a cos  $\omega t$  in the other, there would be no way to make the sum of the two terms to be zero for all time, and we would not have solved the equation. However because the  $(\sin \omega t)$  cancels, we are left with

$$-\omega^2 + \frac{k}{m} \stackrel{?}{=} 0 \tag{18}$$

Equation 18 is easily solved with the choice

$$\omega = \sqrt{\frac{k}{m}} \qquad angular frequency of oscillating cart} \qquad (19)$$

Thus we have shown not only that  $(x = \sin \omega t)$  is a solution of Newton's second law, but we have also solved for the frequency of oscillation  $\omega$ . Newton's second law predicts that the air cart will oscillate at a frequency  $\omega = \sqrt{k/m}$ . This result is easily tested by experiment.

# Exercise 6

(a) In the formula (x = sin  $\omega$ t),  $\omega$  is the angular frequency of oscillation, measured in radians per second. Using the formula  $\omega = \sqrt{k/m}$  and dimensional analysis, find the predicted formula for the period T of the oscillation, the number of *seconds per cycle*.

(b) A mass m = 245 gms is suspended from a spring as shown in Figure (12). The mass is observed to oscillate up and down with a period of 1.37 seconds. From this determine the spring constant k. How could this result have been used in the spring pendulum experiment discussed in Chapter 9 (page 9-3)?

(You can check your answer, since the ball and spring of Figure (12) are the same ones we used in the spring pendulum experiment.)



Figure 12

The spring constant can be determined by measuring the period of oscillation.

The guess we made in Equation 12,  $(x = \sin \omega t)$  is not the only possible solution to our differential Equation 11. In the following exercises, you show that

$$\mathbf{x} = \mathbf{A}\sin\left(\omega t\right) \tag{12a}$$

is also a solution, where A is an arbitrary constant. Since the function  $\sin(\omega t)$  oscillates back and forth between the values + 1 and – 1, the function A sin ( $\omega t$ ) oscillates back and forth between + A and – A. Thus A represents the *amplitude* of the oscillation. The fact that Equation 12a, with arbitrary A, is also a solution to Newton's second law, means that a sine wave with any amplitude is a solution. (This is true as long as you do not stretch the spring too much. If you pull a spring out too far, if you exceed what is called the *elastic limit*, the spring does not return to its original shape and its spring constant changes.)

# **Exercise 7**

As a guess, try Equation 12a as a solution to the differential Equation 11. Follow the same kind of steps we used in checking the guess  $x = sin(\omega t)$ , and see why Equation 12a is a solution for any value of A.



**Figure 13** The exponential decay curve  $e^{-\alpha t}$ .

## **Exercise 8**

(a) Show that the guess

$$x = A\cos(\omega t) \tag{12b}$$

is also a solution to our differential Equation 11. This should be an expected result, because the only difference between a sine wave and a cosines wave is the choice of the time t = 0 when we start measuring the oscillation.

(b) The sine and cosine waves are only special cases of the more general solution

$$x = A \sin(\omega t + \phi)$$
(12c)

where  $\phi$  is the phase of the oscillation discussed in Figure (9). Show that Equation 12c is also a solution of our differential Equation 11. [Hint: the derivative of  $\sin(\omega t + \phi)$  is  $-\omega \cos(\omega t + \phi)$ . You can, if you want, prove this result using Equations 13 and 14 and the trigonometric identities

sin(a+b) = sin a cos b + cos a sin b

 $\cos(a+b) = \cos a \cos b - \sin a \sin b$ 

Remember that  $\phi$  is a constant.]

## **Exercise 9**

We do not want you to think every function is a solution to Equation 11. Try as a guess

$$x = e^{-\alpha t}$$
(20)

which represents an exponentially decaying curve shown in Figure (13). To do this you need to know that

$$\frac{d}{dt}e^{-\alpha t} = -\alpha e^{-\alpha t}$$
(21)

When you try Equation 20 as a guess, what goes wrong? Why can't this be a solution to our differential equation? [Or, by what crazy way could you make it a solution?]

# **Conservation of Energy**

Back in Chapter 10 we calculated the formula for the potential energy stored in the springs when we pulled the cart of Figure (10) a distance x from equilibrium. The result was

spring  
potential 
$$U_{spring} = \frac{1}{2} k x^2$$
 (10-28)

We then used the law of conservation of energy to predict how fast the cart would be moving when it crossed the x = 0 equilibrium line if it were released from rest at a position  $x = x_0$ . The idea was that the potential energy  $\frac{1}{2}kx_0^2$  the springs have when the cart is released, is converted to kinetic energy  $\frac{1}{2}mv_0^2$  the cart has when it is at x = 0 and its speed is v = v\_0.

## Exercise 10

See if you can derive Equation 10-28 without looking back at Chapter 10. If you cannot, review the derivation now.

Using conservation of energy to predict the speed of the air cart was particularly useful back in Chapter 10 because at that time we did not have the analytic solution for the motion of the cart. Now that we have solved Newton's second law to predict the motion of the cart, we can turn the problem around, and see if energy is conserved by the analytic solution.

An analytic solution for the position x(t) and velocity v(t) of the cart is

$$x(t) = \sin(\omega t)$$
  

$$v(t) = \frac{dx}{dt} = \omega \cos(\omega t)$$
(22)

For this solution, the kinetic and potential energies are

kinetic  
energy 
$$\frac{1}{2}$$
 mv<sup>2</sup> =  $\frac{1}{2}$  m $\omega^2 \cos^2(\omega t)$  (23)

potential 
$$\frac{1}{2} kx^2 = \frac{1}{2} k \sin^2(\omega t)$$
 (24)

The total energy  $E_{tot}$  of the cart at any time t is

$$\begin{array}{l} \text{total} \\ \text{energy} &= \begin{array}{c} \text{kinetic} \\ \text{energy} \end{array} + \begin{array}{c} \text{potential} \\ \text{energy} \end{array} \\ \text{E}_{\text{tot}} &= \begin{array}{c} \frac{1}{2} \ \text{m} \ \omega^2 \cos^2(\ \omega t) \end{array} + \begin{array}{c} \frac{1}{2} \ \text{k} \ \sin^2(\ \omega t) \end{array}$$
(25)

At first sight, Equation 25 does not look too promising. It seems that  $E_{tot}$  is some rather complex function of time, hardly what we expect if energy is conserved. However remember that the frequency  $\omega$  is related to the spring constant k by  $\omega = \sqrt{k/m}$ , thus we have

$$\frac{1}{2} m\omega^2 = \frac{1}{2} m \left(\sqrt{\frac{k}{m}}\right)^2$$
$$= \frac{1}{2} m \frac{k}{m} = \frac{1}{2} k$$
(26)

Thus the two terms in our formula for  $E_{tot}$  have the same coefficient  $\frac{1}{2}k$ , and  $E_{tot}$  becomes (using Equation 26 in 25)

$$E_{tot} = \frac{1}{2} k \left[ \cos^2(\omega t) + \sin^2(\omega t) \right]$$
(27)

Equation 27 can be simplified further using the trigonometric identity

$$\cos^{2}(a) + \sin^{2}(a) = 1$$
 (28)

for any value of a. Thus the term in square brackets in Equation 27 has the value 1, and we are left with

$$E_{tot} = \frac{k}{2}$$
(29)

The total energy of the mass and spring system is constant as the oscillator moves back and forth. Energy is conserved after all !

# Exercise 11

What is the total energy of an oscillating mass whose amplitude of oscillation is A? [Start with the solution  $x(t) = A \sin(\omega t)$ , calculate v(t), and then calculate

$$E_{tot} = \frac{1}{2}kx^2 + \frac{1}{2}mv^2$$
.

# THE HARMONIC OSCILLATOR

The sinusoidal motion we have been discussing, which results when an object is subject to a linear restoring force F = -kx, is called *simple harmonic motion* and the oscillating system is often called a *harmonic oscillator*. These general names are used because there are many examples in physics of simple harmonic motion. In some cases the sine wave solution  $sin(\omega t)$  is an exact solution of a Hooke's law problem. In many other cases, the solution is approximate, valid only for small amplitude oscillations where the displacements x do not become too big. In the following sections we will consider examples of both kinds of problems.



Figure 14a

Side view of the torsion pendulum used in the Cavendish experiment.



### Figure 14b

Top view of the torsion pendulum. The light drawing shows the equilibrium position of the pendulum, the dark drawing shows the pendulum displaced by an angle  $\theta$ . In this displaced position, the glass fiber exerts a restoring torque  $\tau_{restoring} = -k\theta$ .

# **The Torsion Pendulum**

One example of simple harmonic motion is provided by part of the apparatus used by Cavendish to detect the gravitational force between two lead balls. The apparatus, illustrated back in Figure (8-8), contains two small lead balls mounted on a light rod, which in turn is suspended from a glass fiber as shown in Figure (14a). (Such glass fibers are easy to make. Heat the center of a glass rod in a Bunsen burner until the glass is about to melt, and then pull the ends of the rod apart. The soft glass stretches out into a long thin fiber.)

If you let the rod with two balls come to rest at its equilibrium position, then rotate then rod by an angle  $\theta$  in the horizontal plane as shown in the top view of Figure (14b), the glass fiber exerts a torque tending to rotate the rod back to its equilibrium position. Careful experiments have shown that the restoring torque exerted by the glass fiber is proportional to the angular distance  $\theta$  that the rod has been rotated from equilibrium. The rod is acting like an angular spring, producing a restoring torque  $\tau_r$  given by an angular version of Hooke's law

$$\tau_{\rm r} = -k\theta \qquad \begin{array}{c} angular \ version \\ of \ Hooke's \ law \end{array} \tag{30}$$

In the Cavendish experiment two large balls of mass M are placed near the small balls as shown in Figure (15).





In the Cavendish experiment, a torque is exerted on the torsion pendulum by the gravitational force of the large lead spheres. In the new equilibrium position, the gravitational torque just balances the restoring torque of the torsion pendulum. The gravitational forces  $\vec{F}_g$  between the large and small balls produce a net torque  $\tau_g$  on the suspended rod of magnitude

$$\tau_{\rm g} = 2 \left( r F_{\rm g} \right) \tag{31}$$

where r is the distance from the center of the rod to the small mass (the lever arm), and the factor of 2 is from the fact that we have a gravitational force acting on each pair of the balls.

The suspended rod will finally come to rest at an angle  $\theta_0$  where the gravitational torque  $\tau_g$  just balances the glass fiber restoring torque  $\tau_r$ , so that there is no net torque the rod. Equating the magnitudes of  $\tau_r$  and  $\tau_g$ , Equations 30 and 31 gives us

$$\begin{aligned} \left| \tau_{g} \right| &= \left| \tau_{r} \right| \\ 2r F_{g} &= k \theta_{0} \end{aligned} \tag{32}$$

Equation 32 can then be solved for the gravitational force  $\tau_g$  in terms of the rod length 2r, the restoring constant k, and the rest angle  $\theta_0$ .

The problem the Cavendish experiment has to overcome is the fact that the gravitational force between the two lead balls is extremely weak. You need an apparatus where the tiny gravitational torque  $\tau_g$  produces an observable deflection  $\theta_0$ . That means that the restoring torque  $\tau_r$  must also be very small. That was why the long glass fiber was used to suspend the rod, for it produces an almost immeasurably small restoring torque.

In order to carry out the experiment and measure the gravitational force  $F_g$ , you need to know the restoring torque constant k that appears in Equation 32. But the feature of the glass fiber that makes it good for the experiment, the small value of k, makes it hard to

directly measure the value of k. To determine k by direct measurement would mean applying known forces of magnitude  $F_g$ , but the only forces around that are sufficiently weak are the gravitational forces you are trying to measure.

Fortunately there is an easy way to obtain an accurate value of the restoring constant k. Remove the large lead balls, displace the rod from equilibrium by some reasonable angle  $\theta$  as shown in Figure (14b), and let go. You will observe the rod to swing back and forth in an oscillatory motion. The rod, two balls, and glass fiber of Figure (14a) form what is called a *torsion pendulum*, and the oscillation is caused by the restoring torque of the glass fiber. The glass fiber is acting like an angular spring, creating an angular harmonic motion in strict analogy to the linear harmonic motion of a mass suspended from a spring.

The analogy applies directly to the equations of motion of the two systems. For a linear one dimensional system like a mass on a spring, Newton's second law is

$$F_x = ma_x = m \frac{d^2x}{dt^2}$$

The angular version of Newton's second law, applied to the simple case of an object rotating about a fixed axis, is from Equation 30 of Chapter 12

$$\tau = I \alpha = I \frac{d^2 \theta}{dt^2}$$
(12-20)

where  $\tau$  is the net torque, I the angular mass or moment of inertia, and  $\alpha$  the angular acceleration of the object about its axis of rotation. For the linear harmonic oscillator (mass on a spring), the force  $F_x$  is a linear restoring force  $F_x = -kx$ , which gives rise to the equation of motion and differential equation

$$F_x = -kx = m \frac{d^2 x}{dt^2}$$
(11a)

$$\frac{\mathrm{d}^2 x}{\mathrm{d}t^2} + \frac{\mathrm{k}}{\mathrm{m}} x = 0 \tag{11}$$

For our torsion pendulum, the restoring torque is  $\tau_r = -k\theta$  which gives rise to the equation of motion and differential equation

$$\tau_{\rm r} = -k\theta = I \frac{d^2\theta}{dt^2}$$
(33a)

$$\frac{d^2\theta}{dt^2} + \frac{k}{I}\theta = 0$$
(33)

Equations 11 and 33 are the same if we substitute the angular distance  $\theta$  for the linear distance x, and the angular mass I for the linear mass m. For the linear motion, we saw that the spring oscillated back and forth at an oscillation frequency  $\omega_0$  and period T given by

$$\omega_0 = \sqrt{\frac{k}{m}}$$
$$T = \frac{2\pi}{\omega_0} = 2\pi \sqrt{\frac{m}{k}}$$
(19)

By strict analogy, we expect the torsion pendulum to oscillate with a frequency  $\omega_0$  and period T given by

$$\omega_0 = \sqrt{\frac{k}{I}}$$
$$T = \frac{2\pi}{\omega_0} = 2\pi \sqrt{\frac{I}{k}}$$
(34)

where I is the moment of inertia of the rod and two balls about the axis defined by the glass fiber (as shown in Figure 14). As a result, by observing the period of oscillation of the rod and two balls (with the big masses M removed), you can determine the restoring constant k of the glass fiber, and use that result in Equation 32 to solve for the gravitational force  $F_g$ . Because you can measure periods accurately by timing many swings, k can be measured accurately, and the Cavendish experiment allows you to do a reasonably good job of measuring the gravitational force  $F_g$ .

## Exercise 12

Solve the differential Equation 33 by starting with the guess

$$\theta(t) = A \sin(\omega_0 t) \tag{35}$$

Check that Equation 35 is in fact a solution of Equation 33, and find the formula for the frequency  $\omega_0$  of the oscillation. Also use dimensional analysis find the period of oscillation.

## **Exercise 13**

In the commercial Cavendish experiment apparatus shown in Figure 8-8, the small balls each have a mass of 170 gms, the distance between the small balls is 12 cm, and the observed period of oscillation is 24 minutes.

(a) Calculate the value of the restoring constant k of the glass fiber.

(b) How big a torque, measured in dyne centimeters, is required to rotate the glass fiber by an angle of one degree. (Remember to convert degrees to radians.)

# The Simple Pendulum

Perhaps the most well-known example of oscillatory motion is the simple pendulum which consists of a mass swinging back and forth on the end of a string or rod. The regular swings of this pendulum serve as the basic timing device of the grandfather's clock.

When we begin to analyze the simple pendulum, we will find that it is not quite so simple after all. The restoring force is not strictly a linear restoring force and we end up with a differential equation whose solution is more complex than the sinusoidal oscillations we have been discussing. What allows us to include this example in our discussion of simple harmonic motion is the fact that, *for small amplitude oscillations*, the restoring force is approximately linear, and the resulting motion is approximately sinusoidal.

Figure (16) is a sketch of a simple pendulum consisting of a small mass m swinging on the end of a string of length  $\ell$ . The downward gravitational force mg has a component of magnitude (mg sin $\theta$ ) directed along the circular path of the ball.

Since the ball is constrained to move along the circular arc, we can analyze the motion of the ball by equating the tangential forces acting along the arc to the mass



## Figure 16

Simple pendulum consisting of a mass swinging on the end of a string. The gravitational force has a component  $mg \sin \theta$ in the direction of motion of the mass.

times the tangential acceleration. The tangential component of the gravitational force is always directed toward the bottom equilibrium position, thus it is a restoring force of the form

$$F_{\text{tangential}} = -\operatorname{mg}\sin\theta \tag{36}$$

As the mass moves along the arc, the speed of the ball is related to the angle  $\theta$  by

$$v_{\text{tangential}} = \ell \frac{d\theta}{dt}$$
 (37)

a result from the beginning of our discussion of circular motion in Chapter 12. (See the discussion before Equation 12-11.) Differentiating Equation 37, we get for the tangential acceleration

$$a_{\text{tangential}} = \frac{dv_{\text{tangential}}}{dt} = \ell \frac{d^2 \theta}{dt^2}$$
 (38)

Thus Newton's second law gives

 $F_{tangential} = ma_{tangential}$ 

$$-\operatorname{mg}\sin\theta = \operatorname{m}\ell \frac{\mathrm{d}^2\theta}{\mathrm{d}t^2}$$
(39)

Dividing through by  $m\ell$  and rearranging terms gives us the differential equation

$$\frac{d^2\theta}{dt^2} + \frac{g}{\ell}\sin\theta = 0 \qquad \begin{array}{c} equation for \ a\\ simple \ pendulum \end{array} (40)$$

Equation 40, the differential equation for the simple pendulum, is more complex than the equations we have been discussing that lead to simple harmonic motion. If you try as a guess that the motion is sinusoidal and try the solution  $\theta = \sin(\omega t)$ , it does not work. You are asked to see why in the following exercise.

## Exercise 14

Try substituting the guess

 $\theta = \sin(\omega t)$ 

into Equation 40 and see what goes wrong. Why can't you make the left side zero with this guess?

There is a solution to Equation 40, it is just not the sine curve we have been discussing. The solution is a curve called an *elliptic integral*, a curve generated much as we generated the sine curve in Figure (3), except that the stick whose shadow generates the curve has to move around an elliptical path rather than around the circular path used in Figure (3). Elliptic integrals carry us farther into the theory of functions than we want to go in this text, thus we will not discuss the exact solution of the differential Equation 40.

# **Small Oscillations**

The problem with Equation 40 is the appearance of the function  $\sin \theta$  in the second term on the left hand side. It is this term that seems to keep us from using the oscillatory solution.

In Figure (17) we look again at the geometry of the simple pendulum. In that figure we have a right triangle whose small angle is  $\theta$ , hypotenuse the string length  $\ell$ , and opposite side x. The definition of the sine of the angle  $\theta$  is

$$\sin \theta \equiv \frac{\text{opposite side}}{\text{hypotenuse}} = \frac{x}{\ell}$$
(41)



arc length



$$\sin \theta = \frac{x}{\ell} \qquad \theta = \frac{x}{\ell}$$

The definition of the angle  $\theta$ , in radian measure, is the arc length divided by the radius  $\ell$  of the circular arc

$$\theta \equiv \frac{\text{arc length}}{\text{radius}} = \frac{\text{arc length}}{\ell}$$
(42)

From Figure (17) we see that for small angles  $\theta$  the opposite side x and the arc length l are about the same. The smaller the angle  $\theta$ , the more nearly equal they are. If we restrict our analysis to small amplitude swings, we can replace sin  $\theta$  by  $\theta$  in Equation 40, giving us the differential equation

$$\frac{d^2\theta}{dt^2} + \frac{g}{\ell}\theta = 0$$
equation for small  
oscillations of a  
simple pendulum
(43)

Equation 43 is an equation for simple harmonic motion. If we try the guess  $\theta = \sin(\omega_0 t)$ , and plug the guess into Equation 43, we can solve the equation provided the frequency  $\omega_0$  and the period of oscillation T have the values

$$\omega_{0} = \sqrt{\frac{g}{\ell}}$$
$$T = \frac{2\pi}{\omega_{0}} = 2\pi \sqrt{\frac{\ell}{g}} \qquad period of a \\simple pendulum \qquad (44)$$

## Exercise 15

Substitute the guess  $\theta = A \sin(\omega_0 t)$  into Equation 43 and show that you get a solution provided  $\omega_0^2 = g/\ell$ . Then use dimensional analysis to derive a formula for the period of the oscillation.

From Equation 44, we see that the period of the oscillation of a simple pendulum depends only on the gravitational acceleration g and the length  $\ell$  of the pendulum. It does not depend on!tXe mass m of the swinging object, nor on the amplitude of the oscillation, provided that the amplitude is kept small. For these reasons the simple pendulum makes a good timing device.

## **Exercise 16**

How long should a simple pendulum be so that it's period of oscillation is one second?

# Simple and Conical Pendulums

In Chapter 9 we analyzed the motion of a conical pendulum. The conical pendulum also consists of a mass on a string, but the mass is swung around in a circle as shown in Figure (18), rather than back and forth along an arc as for a simple pendulum.

From our analysis of the conical pendulum, we found that the period of rotation was given by the formula

$$T_{cp} = 2\pi \sqrt{\frac{h}{g}}$$
 period of a (9-34)

where h is the height shown in Figure (18). Considering the trouble we went through to get an approximate solution to the simple pendulum, it seems surprising that Equation 9-34 is an exact solution to Newton's second law for any achievable radius x of the circle.

For small circles, where  $x \ll \ell$ , the height h and the string length  $\ell$  are approximately the same and we have

$$T_{cp} = 2\pi \sqrt{\frac{g}{h}} \approx 2\pi \sqrt{\frac{g}{\ell}}$$
 (45)

But this is just the period of a simple pendulum if the oscillations are kept small. Since the two pendulums have the same period for small oscillations, it makes no difference, as far as the period is concerned, whether we swing the balls back and forth or around in a circle. This prediction is easily checked by experiment.



**Figure 18** *The conical pendulum.* 

## Exercise 17

You can do your own experiments to show that as you increase the amplitude of a simple pendulum, the period of oscillation starts to get longer. In contrast, when you increase the radius of the circle for a conical pendulum, the height h and the period  $T_{cp}$  become shorter.

(a) From your own experiments estimate how much longer the period of a simple pendulum is when the maximum angle  $\theta_{max}$  is 90° than when  $\theta_{max}$  is small. (Is it 20% longer, 30% longer? Do the experiment and find out. Does this percentage depend on the length  $\ell$  of the string?

(b) For a conical pendulum, at what angle  $\theta_0$  (shown in Figure 18) is the period half as long as it is for small angles  $\theta_0$ ? Give your answer for  $\theta_0$  in degrees.

# Exercise 18

Another way to analyze the simple pendulum is to treat the mass and the string as a rigid object that can rotate about an axis through the top end of the string as shown in Figure (19). Then use the angular version of Newton's second law in the form

$$\tau = I \frac{d^2 \theta}{dt^2}$$
(12-20)

where  $\tau$  is the net torque, and (I) the moment of inertia, of the mass and string about the axis 0.

(a) When the string is at an  $\theta$  angle as shown in Figure (19), what is the torque  $\tau$  about the axis 0, exerted by the gravitational force mg?

(b) What is the moment of inertia I of the mass and string about the axis 0?

(c) Show that when you use the above values for  $\tau$  and (I) in Equation 12-20 you get the same differential Equation 40 that we got earlier for the simple pendulum.

# Exercise 19 A Physical Pendulum

A uniform rod of length  $\ell$  is pivoted at one end as shown in Figure (20). It is free to swing back and forth about this axis, forming what is called a physical pendulum. A *simple pendulum* is one where the mass is all concentrated at the end as in Figure (19). In a *physical pendulum* the mass is distributed in some other way, in this case uniformly along the rod.

(a) What is the torque  $\tau$  about the axis 0 exerted by the gravitational force on the rod? (In Chapter 13 near Equation 13-11, we showed that when calculating the torque exerted by a gravitational force, you may assume that all the mass is concentrated at the center of gravity of the object.)

(b) What is the moment of inertia (I) of the rod about the axis at the end of the rod? (See exercise 5 in Chapter 12.)

- (c) Write the differential equation for the motion of the rod. (Use the procedure outlined in Exercise 18.)
- (d) Find the period of small oscillations of the rod.







**Figure 20** *A physical pendulum.* 

# NON LINEAR RESTORING FORCES

The simple pendulum is an example of an oscillator with a non linear restoring force. In Figure (21), we show the actual restoring force (mg sin  $\theta$ ) and the linear approximation (mg  $\theta$ ) that we used in order to solve the differential equation for the pendulum's motion. You can see that if the angle  $\theta$  always remains small, much less than  $\pi/2$  in magnitude, then the linear force (mg  $\theta$ ) is a good approximation to the non linear force (mg sin  $\theta$ ). Since the linear force gives rise to sinusoidal simple harmonic motion, we expect sinusoidal motion for small oscillations of the simple pendulum. What we are seeing is that a linear restoring force is described by a straight line, and that the non linear restoring force can be approximated by a straight line in the region of small oscillations. In physics, there are many examples of complex, non linear restoring forces which for small amplitudes can be approximated by a linear restoring force, and which therefore lead to small amplitude sinusoidal oscillations. A rather wild example which we will discuss shortly, is the collapse of the Tacoma Narrows bridge. The bridge undoubtedly started oscillating with small amplitude sinusoidal oscillations. What happened was that these oscillations were continually driven by the shedding wind vortices until the amplitude of oscillation became large and the restoring force was no longer linear. (There was still a more or less sinusoidal motion almost up to the point when the bridge collapsed.)



## Figure 21

The non linear restoring force  $mg \sin\theta$  can be approximated by the straight line (linear term)  $mg \theta$  if we keep the angle  $\theta$  small.

# **MOLECULAR FORCES**

One of the most important examples of a non linear restoring force is the molecular force between atoms. Consider, for example, the hydrogen molecule which consists of two hydrogen atoms held together by a molecular force. (We will discuss the origin of the molecular force in Chapter 17.)

In the hydrogen molecule, the hydrogen atoms have an equilibrium separation, and the molecular force provides a restoring force to this equilibrium separation. The restoring force, however, is quite non linear. If you try to squeeze the atoms together, you quickly build up a large repulsive force that keeps the atoms from penetrating far into each other.

If you try to pull the atoms apart, there is an attractive force that pulls the atoms back together. The attractive force never gets too big, and then dies out when the separation gets much larger than an atomic diameter.

In Figure (22) we have sketched the molecular force as a function of the separation of the atoms, the origin being at the equilibrium position. This graph is not too unlike Figure (21) where we have the force curve for the simple pendulum. For the pendulum, the equilibrium position is at  $\theta = 0$ , thus the origin of both curves represents the equilibrium position.



Figure 22

Sketch of the molecular force between two hydrogen atoms. As long as the atoms stay close to the equilibrium position, the force can be represented by a straight line—a linear restoring force.

While the overall shape of the force curves for the simple pendulum and the molecular force are quite different, right in close to the origin both curves can be approximated by a straight line, a linear restoring force. As long as the amplitudes of the oscillation remain small, we effectively have a linear restoring force and any oscillations should be simple harmonic motion.

In Chemistry texts one often sees molecular forces as being represented by springs as shown in Figure (23). The spring force, given by Hooke's law, is our ideal example of a linear restoring force. We can now see that, while the molecular force in Figure (22) does not look like a linear spring force, if the amplitude of oscillation remains small, the spring force provides a reasonably good approximation to the actual molecular force. The chemist's diagrams are not so bad after all.

In a crystal, like quartz, where you have many atoms held together by molecular forces, it is possible to get all the atoms oscillating together. Each atom only oscillates a very small distance about its equilibrium position, but all the oscillations can add up to produce a fairly large, quite detectable oscillation of the crystal as a whole. An advantage of a quartz crystal is that these oscillations can be both driven and detected by electric fields. This vibration or simple harmonic motion of a small quartz crystal is used as the basic timing device for digital watches, computers, and almost all forms of modern electronics.

In Galileo's time we used small oscillations of a non linear harmonic oscillator, the simple pendulum, as a basic time device. Now we use the small oscillations of a non linear harmonic oscillator, the atoms in a quartz crystal, as our most convenient timing device. The main thing we have changed in the last 300 years is not the basic physics, but the size and frequency of the device.



**Figure 23** *Representation of the molecular force by a spring force.* 

# DAMPED HARMONIC MOTION

If you start a pendulum swinging or a quartz crystal oscillating, and do not keep the oscillation going with some kind of external force, the oscillation will eventually die out due to friction forces. Such a dying oscillation is called *damped harmonic motion*.

The analysis of damped harmonic motion starts out quite easily. In Newton's second law add a damping force like the air resistance term we added to our analysis of projectile motion. (See Chapter 3, Figure 31.) We could write, for example

$$F_{tot} = F_{restoring} + F_{damping}$$

$$F_{tot} = -kx - bv_x$$
(46)

where x(t) is the coordinate of the oscillator,  $v_x(t) = dx/dt$  its velocity, and we are assuming a simple linear damping proportional to (-v) with a strength b.

Using Equation 46 in Newton's second law gives

$$F_{tot} = -kx - bv_x = ma_x \tag{47}$$

With  $a_x = d^2 x/dt^2$ , this becomes, after dividing through by m and rearranging terms

$$\frac{d^2x}{dt^2} + \frac{b}{m}\frac{dx}{dt} + \frac{k}{m}x = 0$$
(48)

Equation 48 is our new differential equation for damped harmonic motion. It is like our old differential Equation 11a for undamped oscillation, except that it has the additional term  $\frac{b}{m}\frac{dx}{dt}$  representing the damping.

If we were doing a computer solution of harmonic motion, adding the damping term represents hardly any extra effort at all. In the appendix to this chapter we discuss a short computer program to handle harmonic motion. Starting with the first version of the program that has no damping, you can include damping by changing the line

LET 
$$F = -k * x$$

to the new line

$$\text{LET F} = -\mathbf{k} * \mathbf{x} - \mathbf{b} * \mathbf{v} \tag{49}$$

and placing your choice for the damping constant b in the initial conditions.

In contrast, when working with differential equations analytically you find that a very small change in the equation can make a great deal of difference in the effort required to obtain a solution. Adding a bit of damping to a harmonic oscillator changes the curve from a pure sinusoidal motion to a dying sine wave. If you try using a pure sine wave as a guess for the solution to the differential Equation 48 for damped harmonic motion, the guess does not work because the pure sine wave has the wrong shape. The decay of the sine wave has to be built into your guess before the guess stands a chance of working.

The difficult part about solving differential equations is that you essentially have to know the answer before you can solve the equation. You only have to know general features like the fact that in working with Equation 11a you are dealing with a sinusoidal oscillation. You can then use the differential equation to determine explicit features like the frequency of the oscillation. It is helpful to have a physical example to tell you what the general features of the motion are, so that you can begin the process of solving the equation. That is why we begin this chapter with the demonstration in Figure (2) that the motion of a mass on a spring is similar to circular motion seen sideways, namely sinusoidal motion. To set up a physical model for damped harmonic motion is not too difficult. One way to add damping to the air cart and springs oscillator is to run a string from the air cart over a pulley to a small weight as shown in Figure (24). The idea was to have the weight move up and down in a glass of water to give us fluid damping. But it turned out that there was enough friction in the pulley itself to give us considerable damping.

To record the motion of the cart, we used the air cart velocity detector that we used in Chapter 8 to study the momentum of air carts during collisions. Figure (25a) shows the velocity of the air cart damped only by the friction in the pulley. In Figure (25b) water was added to the glass so that the weight on the string was moving up and down in water. The result was considerably more damping with the curve almost dying out before any oscillations take place.

It turns out that mechanical oscillators like a pendulum or a mass on a spring are not particularly convenient devices for studying damped harmonic motion, or forced harmonic motion which is the subject of the next



Adding damping to the air cart oscillator.



**Figure 25a** Damping caused by the pulley and weight alone.

section. It is hard to control the damping, just adding the pulley in Figure (24) gave us almost too much damping. Worse yet, the damping that we get from friction in a pulley, or a mass moving up and down in water, is not a simple linear damping force of the form – bv. What is remarkable about these systems is that much more complex forms of damping give us results similar to what we would get with linear damping.

In Chapter 27 we will study the behavior of basic electric circuits made from electrical components called capacitors, inductors, and resistors. It turns out that the amplitude of the currents in these circuits obey differential equations that are exactly like our oscillator Equations (11) and (45). The damping is caused by the resistor in the circuit, the damping is accurately given by a linear damping term proportional to the amount of resistance in the circuit. (The resistance can be changed simply by turning a knob on a resistance pot.)

Figure (26), taken from Chapter 27, is an example of damped harmonic motion in an electric circuit. Here we have a curve with enough oscillations so that we can see how the wave is damped. In Chapter 27 we will see that the amplitude of the oscillation dies exponentially, following a mathematical curve of the form  $e^{-\alpha t}$ . As a result, the wave in Figure (26) has the form

$$x = (Ae^{-\alpha t}) \sin \omega t$$

$$decaying \quad sine wave \quad (50)$$

$$amplitude \quad oscillation$$



**Figure 25b** *Resulting motion when water was added to the glass.* 

It turns out that if we use Equation 50 as our guess for a solution to our differential Equation 48 for damped harmonic motion, the guess works, and we can determine both the frequency  $\omega$  and decay rate  $\alpha$  in terms of the constants that appear in Equation 48.

When we study electric circuits, you will get much more experience with the exponential function  $e^{-\alpha t}$ , and you will have a better laboratory setup for studying damped and forced harmonic motion. In other words, now, with our somewhat crude mechanical experiments and lack of familiarity with exponential damping, is not the best part of the course to go deeply into the mathematical analysis of these motions. What we will do instead is discuss the motions more or less qualitatively and leave the more detailed analysis for later.

### **Exercise 20 Damped Harmonic Motion**

We won't let you completely off the hook for doing mathematical analysis of damped harmonic motion. Start with Equation 5a as a guess for the form for the displacement x(t) for a damped harmonic oscillator

$$x(t) = Ae^{-\alpha t} \sin(\omega t)$$



#### Figure 26

Damped harmonic motion seen in an electric circuit. Note the difference in time scales. The electrical oscillations we will study are usually of much higher frequency than the mechanical ones.

use the following rules of differentiation to calculate dx/dt and  $d^2x/dt^2$ 

$$\frac{d}{dt}e^{-\alpha t} = -\alpha e^{-\alpha t}$$
$$\frac{d}{dt}[a(t)b(t)] = \frac{da}{dt}b + a\frac{db}{dt}$$

and show that when you try this guess in the differential Equation 48, you do in fact get a solution, and that  $\omega$  and  $\alpha$  are given by

$$\omega = \sqrt{\frac{k}{m} - \frac{b^2}{4m^2}} \qquad \alpha = \frac{b}{2m}$$
(51)

You can see that in the absence of damping, when b = 0, we get back to our old result  $\omega$  =  $\sqrt{k/m}$  .

## **Critical Damping**

In Figure (25b) the damping was so great that the motion damped out almost before the curve had a chance to oscillate. It turns out that there is a critical amount of damping that just kills all oscillations. Any further increase in damping and the mass just coasts to rest.

The idea of critical damping can be seen in our analytic solution for damped harmonic motion obtained Exercise 20. Equation 51 gives us a formula for the frequency of oscillation  $\omega$  in terms of the constants k and b. We can see as the frequency of oscillation goes to zero, i.e., the period of oscillation becomes infinite when

$$\frac{k}{m} = \frac{b^2}{4m^2}; \quad b = 2\sqrt{mk}$$
$$b = \sqrt{4mk} \qquad \begin{array}{c} critical\\ damping \end{array}$$
(52)

Equation 52 is the condition for critical damping because if the period of oscillation is infinite there are no oscillations.

# RESONANCE

When you are pushing a child on a swing, you time your pushes to coincide with the motion of the child. Usually you give a shove just after the child has swung back and is starting forward again.

If you push forward just as the child is swinging forward, your force  $\vec{F}$  and the child's velocity  $\vec{v}$  are in the same direction, the dot product  $\vec{F} \cdot \vec{v}$  is positive, and you are adding energy to the child's motion. Initially the energy you add goes into increasing the amplitude of the swing. After a while friction effects become large enough that the energy you add in each push is dissipated by friction in each swing. (If there is not enough friction, or you push too hard, the child will end up going over the top.)

The key to getting the child swinging was to time your shoves so that  $\vec{F} \cdot \vec{v}$  was always positive. If you pushed the child at random intervals, so that  $\vec{F} \cdot \vec{v}$  was sometimes positive, sometimes negative, you would be sometimes adding energy and sometimes removing it. The net result would be that your shoves would not be particularly effective in helping the child to swing.

To make sure that you are always adding energy to the child's swing, you want to time your shoves with the natural frequency of oscillation of the child. When you do this, we say that your shoves are in *resonance* with the oscillation of the child.

The striking feature of resonance is that a small repeated force can produce a large oscillation. If the damping is small then by adding just a little energy with each shove, the energy accumulates until you end up with a very energetic oscillation. A rather dramatic consequence of this effect is shown in Figure (27) where we see the Tacoma Narrows bridge oscillating wildly and then collapsing.

The new bridge was dedicated in April of 1940. Three months later a reasonably stiff breeze started the bridge oscillating, an oscillation that finally destroyed the bridge's integrity.

The brute force of the wind itself did not destroy the bridge. The bridge was designed to handle far stronger winds. What happened was that as the wind was blowing over the bridge, vortices began to peel off the bridge. Whenever fluid flows past a cylindrical object at the right speed, vortices began to peel off, first on one side of the cylinder, then the other, and are carried downstream, forming a wake of vortices seen in the wind tunnel photograph of Figure (28). This vortex structure is called a *Karmen vortex street* after the hydrodynamicist Theodore Von Karmen.

In the case of the Tacoma Narrows bridge, vortices alternately peeled off the top and bottom of the down wind side of the bridge, rocking the bridge at its natural frequency of oscillation. While no separate jolt by any one vortex would have much effect on the bridge, the



**Figure 27a** Tacoma Narrows bridge oscillating in the winds of a mild gale on July 1, 1940.



**Figure 27b** After a couple of hours the bridge collapsed.
resonance between the peeling of vortices and the oscillation of the bridge caused the oscillations to grow to destructive proportions.

The example of the Tacoma Narrows bridge illustrates how widely the ideas of simple harmonic motion and resonance apply to physical systems. The bridge is far more complex than a mass on a spring, and the *vortex street* (line of vortices) exerts a rather complex driving force. However the bridge had a natural frequency, the vortices provided a small driving force at that frequency, and we got a resonant amplification of the oscillation.

To apply Newton's second law to resonant motion, we have to add an oscillating driving force to the system under study. As we have seen from the Tacoma Narrows bridge discussion, we do not need to know the exact form of the driving force, all we need is a repetitive force that can be timed with the natural oscillation. For the theoretical analysis we can use the simplest mathematical form we can find for the driving force, which turns out to be a sine wave.

To write a formula for the driving force, let  $\omega_0$  be the natural frequency of oscillation ( $\omega_0 = \sqrt{k/m}$  if the damping is small), and let  $\omega$  be the frequency of the driving force. Then the total force, acting on the oscillating system like a mass on a spring, can be written

$$F_{x \text{ tot}} = -kx - bv + F_d \sin(\omega t)$$
 (52)



#### Figure 28

Karman vortex street in the flow of water past a circular cylinder. The vortices peel off of alternate sides of the cylinder and flow downstream forming a double line of vortices. (Reynolds number = 140.) Photograph by Sadatoshi Taneda.

where in the driving term,  $F_d$ , represents the amplitude or strength of the sinusoidal driving force. Using Equation 52 in Newton's second law  $F_{x \text{ tot}} = ma_x$  gives

$$-kx - b\frac{dx}{dt} + F_d \sin\left(\omega t\right) = m\frac{d^2x}{dt^2}$$
(53)

Dividing through by m and rearranging terms gives

$$\frac{d^{2}x}{dt^{2}} + \frac{b}{m}\frac{dx}{dt} + \frac{kx}{m} = F_{d}\sin\left(\omega t\right)$$
(54)

Equation 54 is the standard form for the differential equation representing forced or resonant harmonic motion. It is the simplest equation we can write whose solution has the features we associate with the phenomena of resonance.

In our study of electric circuits, we can easily create a circuit whose behavior is accurately described by Equation 54. We saw that we could use resistors to add linear damping of the form – bv. It is not hard to add a purely sinusoidal driving force of the form  $F_d \sin (\omega t)$ , where we can adjust the driving frequency by turning a knob. In other words, with electric circuits we can accurately study the predictions of Equation 54.

With mechanical systems like a mass on a spring, it is hard to get linear damping, and the sinusoidal driving force is usually simulated by some trick such as wiggling the supported end of the spring at a frequency  $\omega$ . Despite the crudeness of the experiment, the equation gives a surprisingly good prediction of what we see.

Since we will later have a laboratory setup that accurately matches Equation 54, we will postpone (until Appendix 1) the mathematical solution of Equation 54. Instead we will investigate the resonance phenomena qualitatively, using the simple setup of a mass on a spring, where we hold the other end of the spring in our hand and move our hand up and down at a frequency  $\omega$  as shown in Figure (29).

## **Resonance Phenomena**

We have already seen that if we hold our hand still, pull the mass down, and let go, the mass oscillates up and down at the natural frequency  $\omega_0 = \sqrt{k/m}$ . You can observe some damping, because the mass finally stops oscillating. But the damping is small and has no noticeable effect on the resonant frequency. (We can neglect the term  $b^2/4m^2$  in Equation 51.)

Now try a very different experiment. Stop the mass from oscillating, and slowly move your hand up and down a small distance. If you do this slowly enough and carefully enough, the mass will move up and down with your hand (just as if the spring were not there). In this case the formula for the motion of the mass is

$$\mathbf{x} = \mathbf{x}_0 \sin(\omega t) \qquad \boldsymbol{\omega} << \boldsymbol{\omega}_0 \tag{55}$$

where the frequency  $\omega$  of the oscillation of the mass is the frequency of  $\omega$  the oscillation of your hand. This only happens if you oscillate your hand at a frequency much much lower than the natural oscillation frequency  $\omega_0$ .

In the next experiment, keep the oscillations of your hand small in amplitude, but start moving your hand up and down rapidly, at a frequency  $\omega$  considerably greater than the natural frequency  $\omega_0$ . Now, what happens is that the mass oscillates at the same frequency as your hand, but out of phase. When your hand is going down, the mass is coming up, and vice versa. Now the formula for the displacement x of the mass is

$$\mathbf{x} = -\mathbf{x}_0 \sin(\omega t) \qquad \omega >> \omega_0 \tag{56}$$

where the minus sign tells us that the mass is oscillating out of phase with our hand.

A way we can write both Equations 55 and 56 is in the form





$$\mathbf{x} = \mathbf{x}_0 \sin\left(\omega t + \phi\right) \tag{57}$$

where  $\phi$  is the phase angle of the oscillation (see Figures 8 and 9 and Equation 4 at the beginning of this chapter for a discussion of phase angle.) In Equation 56, where  $\omega \ll \omega_0$  and there was no phase difference, the phase angle  $\phi$  is zero. In Equation 56 where  $\omega \gg \omega_0$ , and the motion is completely out of phase, the phase angle  $\phi$  is  $\pi$  or 180°.

Equations 55 and 56 represent the two extremes of driven harmonic motion. The mass moves with a small amplitude at the same frequency  $\omega$  as the driving force. When the driving frequency is much less than the natural frequency  $\omega_0$ , the difference in phase between the driving force and the response of the mass is zero degrees. When  $\omega \gg \omega_0$  the phase difference increases to  $\pi$  or 180°.

As the third experiment, start at the low frequency where the mass is following your hand, and slowly increase the frequency  $\omega$  of oscillation of your hand, keeping the amplitude of oscillation constant. As  $\omega$ approaches  $\omega_0$ , the amplitude of oscillation of the mass increases. When you get close to the natural frequency  $\omega_0$ , the oscillation becomes so large that the mass will most likely jump off the spring. This is the phenomena of resonance, the phenomenon that destroyed the Tacoma Narrows bridge.

How big the oscillation of the mass becomes depends mainly how close you are to resonance, how close  $\omega$  is to  $\omega_0$ , and how big the damping force is relative to the driving force. The formula for the amplitude  $x_0$  of the motion of the mass obtained by solving Equation 54 is

$$x_0 = \frac{F_d/m}{\sqrt{(\omega^2 - \omega_0^2) + (b/m)^2 \omega^2}}$$
 (58)

where  $F_d$  is the strength of the driving force,  $\omega$  the driving frequency,  $\omega_0 = \sqrt{k/m}$  the natural frequency and b the damping constant. In the absence of damping (b=0), Equation 58 predicts an infinite amplitude at the resonant frequency  $\omega = \omega_0$ . Such an infinite amplitude is prevented either by damping or by the destruction of the system. (A damping mechanism could have saved the Tacoma Narrows bridge.)

#### Exercise 21

To derive Equation 21, you start with Equation 57 as a guess

$$x = x_0 \sin(\omega t + \phi)$$
(57)

and substitute that into the differential Equation 54. It turns out that you can get a solution provided that the amplitude  $x_0$  has the value given by Equation 58, and the phase angle  $\phi$  is given by

$$\tan(\phi) = \frac{(b/m)\omega}{\omega^2 - \omega_0^2}$$
(59)

Doing the work, actually substituting Equation 57 into 54 and getting Equations 58 and 59 for  $x_0$  and  $\phi$  is a somewhat messy job which we leave to Appendix 1 on the next page. Here we would rather have you develop an intuitive feeling for the solutions in Equations 58 and 59 by working the following exercises.

(a) Write the formula for  $x_0$  in the case b = 0. Sketch the resulting curve using the axes shown in Figure (30). Explain what happens as  $\omega \to \omega_0$ .

(b) When the damping is not zero, find the formula for the amplitude of oscillation  $x_0$  at the resonant frequency  $\omega\to\omega_0$ . Check the dimensions of your answer.

(c) What is the phase angle  $\phi$  at resonance? How does the phase angle change as we go from  $\omega \ll \omega_0$  to  $\omega \gg \omega_0$ ?

In Figure (31) we have graphed the amplitude  $x_0$  for a fixed driving force  $F_d$ , as a function of  $\omega$  for several values of the damping constant b. The main point to get from this diagram is that the smaller the damping, the sharper the resonance.



#### Transients

There is one more qualitative experiment we want to do with our simple apparatus of the hand held spring and mass of Figure (29). Instead of gently starting the mass moving as we had you do in the earlier experiments, let the mass fall from some small height and move your hand up and down at the same time.

If you just let the mass drop from some small height, it will oscillate up and down at the resonant frequency  $\omega_0$ . If you just start moving your hand slowly at a frequency  $\omega$ , the mass will move at the same frequency as your hand, building up to an amplitude given more or less by Equation 58 and shown in Figure (31), the driven oscillation we have been discussing.

If you drop the weight and move your hand at the same time, you get both kinds of motion at once. You get the natural oscillation at a frequency  $\omega_0$  that eventually dies out due to damping, and the driven oscillation at the frequency  $\omega$  that eventually builds up to an amplitude  $x_0$ . For a while, before the natural oscillation has died out, the resulting motion is a mixture of two frequencies of oscillation and can look quite complex. The natural oscillation is called a *transient* because it eventually dies out. But until the transients do die out, forced harmonic motion can be fairly complicated to analyze. In the next chapter we will study a powerful technique called *Fourier analysis* that allows us to study complex motions that involve such a mixture of oscillations.



Amplitude of the oscillation for various values of the damping constant. The amplitude of the driving force  $F_d$  is the same for all curves.

## APPENDIX 14–1 SOLUTION OF THE DIFFERENTIAL EQUATION FOR FORCED HARMONIC MOTION

The equation we wish to solve is Equation 54

$$\frac{d^2x}{dt^2} + \frac{b}{m}\frac{dx}{dt} + \frac{k}{m}x = \frac{F_d}{m}\sin\left(\omega t\right)$$
(54)

where our guess for a solution is Equation 57

$$\mathbf{x} = \mathbf{x}_0 \sin\left(\omega \mathbf{t} + \mathbf{\phi}\right) \tag{57}$$

The quantities  $x_0$  and  $\phi$  are the unknown amplitude and phase of the oscillation that we wish to determine.

To simply plug our guess into Equation 54 and grind away leads to a sufficiently big mess that we could easily make a mistake. We will instead simplify things as much as possible to make the calculation easier. The first step is to define the constants

$$\omega_0^2 = k/m; b' = b/m; F' = F_d/m$$
 (60)

Next we wish to get the phase angle into the forcing term so that it appears only once in our equation. We can do this by using a time scale t' where

$$\omega t' = \omega t + \phi \implies \omega t = \omega t' - \phi \tag{61}$$

In terms of the new constants and t' our differential equation becomes

$$\frac{d^2x}{dt^2} + b'\frac{dx}{dt} + \omega_0^2 x = F'\sin\left(\omega t' - \phi\right) (62)$$

Our guess, and its first and second derivative are

$$\mathbf{x} = \mathbf{x}_0 \sin\left(\boldsymbol{\omega} t'\right) \tag{63a}$$

$$\frac{\mathrm{d}x}{\mathrm{d}t} = \omega \, \mathbf{x}_0 \cos\left(\omega t'\right) \tag{63b}$$

$$\frac{\mathrm{d}^2 x}{\mathrm{d}t^2} = -\omega^2 x_0 \sin\left(\omega t'\right) \tag{63c}$$

In deriving dx/dt we used the fact that

$$\frac{\mathrm{d}x}{\mathrm{d}t} = \frac{\mathrm{d}x}{\mathrm{d}t'}\frac{\mathrm{d}t'}{\mathrm{d}t} = \frac{\mathrm{d}x}{\mathrm{d}t'}$$

where  $t' = t - \phi/\omega$  so that dt'/dt = 1. We will also use the trigonometric identity

$$\sin (a + b) = \sin (a) \cos (b) + \cos (a) \sin (b)$$

to write

$$F' \sin (\omega t' - \phi)$$
  
= F' sin (\omega t')cos (-\phi) + F' sin (-\phi)cos (\omega t') (64)  
= F' sin (\omega t')cos (\phi) - F' sin (\phi)cos (\omega t')

where we used

$$\cos(-\phi) = \cos(\phi), \sin(-\phi) = -\sin(\phi)$$

Substituting Equation 63 and 64 into 62, and separately collecting terms with  $\sin(\omega t')$  and  $\cos(\omega t')$ , we get

$$\sin (\omega t') \left[ -\omega^2 x_0 + \omega_0^2 x_0 - F' \cos (\phi) \right] + \cos (\omega t') \left[ \omega b' x_0 + F' \sin (\phi) \right]$$
$$= 0$$
(65)

Because there are both  $\cos(\omega t')$  terms and  $\sin(\omega t')$  terms in Equation 65, there is no way to make everything add up to zero for all times unless the coefficients of both  $\cos(\omega t')$  and of  $\sin(\omega t')$  are separately equal to zero. This gives us the two equations

$$F'\sin(\phi) = -\omega b' x_0 \tag{66}$$

$$F'\cos\left(\phi\right) = \left(\omega_0^2 - \omega^2\right) x_0 \tag{67}$$

If we divide Equation 66 by Equation 67, the F' and  $x_0$  cancel and we get

$$\frac{\sin(\phi)}{\cos(\phi)} = \frac{-\omega b'}{\omega_0^2 - \omega^2}$$

$$\tan(\phi) = \frac{(b/m)\omega}{\omega^2 - \omega_0^2}$$
(59)

This is the result we stated earlier, namely Equation 59.

To solve for the amplitude  $x_0$  of the oscillation, we can use Equation 66 to get

$$\mathbf{x}_0 = \frac{\mathbf{F}'}{\boldsymbol{\omega}\mathbf{b}'} \sin\left(\boldsymbol{\phi}\right) \tag{68}$$

To find the  $\sin(\phi)$  from the  $\tan(\phi)$  which we already know, construct the right triangle shown in Figure (32). We have made the opposite and adjacent sides so that the ratio comes out as  $\tan(\phi)$ , and the hypotenuse is given by the Pythagorean theorem. Thus  $\sin(\phi)$  is

$$\sin\left(\phi\right) = \frac{\omega b'}{\sqrt{\left(\omega^2 - \omega_0^2\right) + \omega^2 {b'}^2}} \tag{69}$$

Substituting 69 into 68, and setting b' = b/m,  $F' = F_d/m$ , we get

$$x_{0} = \frac{F_{d}/m}{\sqrt{\left(\omega^{2} - \omega_{0}^{2}\right) + \omega^{2}b^{2}/m^{2}}}$$
(58)

which is our earlier Equation 58 for  $x_0$ .

#### Transients

In our qualitative discussion of forced harmonic motion, we saw that in addition to the driven oscillation  $x_{driven} = x_0 \sin(\omega t + \phi)$  we have just studied, we could also have transient motion at the natural frequency  $\omega_0$ . In controlled experiments, you observe that any transient motion present initially finally dies out and you are eventually left with just the driven motion. The transient motion  $x_{tr}(t)$  is just damped harmonic motion that satisfies the equation of motion

$$\frac{d^2x}{dt^2} + \frac{b}{m}\frac{dx}{dt} + \frac{k}{m}x = 0$$
(48)

The question we wish to answer now is whether the forced harmonic motion equation

$$\frac{d^2x}{dt^2} + \frac{b}{m}\frac{dx}{dt} + \frac{k}{m}x = F_d \sin\left(\omega t\right)$$
(54)

allows us to have both driven and transient motion at the same time. In other words, is a guess of the form

$$\mathbf{x} = \mathbf{x}_0 \sin\left(\omega \mathbf{t} + \mathbf{\phi}\right) + \mathbf{x}_{\mathrm{tr}} \tag{70}$$

where x is the sum of the driven motion and an arbitrary amount of transient motion, is this sum also a solution of Equation 54?

If you substitute our new guess 70 into 54, the driven term satisfies the whole equation and the transient terms add up to zero because of Equation 48, thus we do get a solution. Transient motions are allowed by Newton's second law.



**Figure 32** *Triangle to go from tan \theta to sin \theta.* 

## APPENDIX 14-2 COMPUTER ANALYSIS OF OSCILLATORY MOTION

In this appendix, we will use the computer to analyze the motion of the oscillating air cart shown in Figure (33). For this problem, the computer solution does not have the elegance of the calculus solution we have been discussing. The calculus approach gives us a single solution valid for all values of the experimental parameters. With the computer we have to alter the program and rerun it any time we want to change a parameter such as the mass of the cart, the spring constant, or the initial position or velocity. The calculus approach gives us a single formula valid for all values of the experimental parameters.

However, the advantage of using the computer is that we can easily modify the program to include new physical phenomena. For example, to add damping, all we have to do is change the command

LET 
$$F = -K * X$$

to the command

LET F = -K\*X - b\*V

and rerun the program. To add damping to the calculus solution, we had to work with a differential equation (48) that was much more difficult to solve than the equation for undamped motion (11).

The computer opens up a number of possibilities for student project work. For example, in our discussion of the simple pendulum shown in Figure (21), we had to limit our analysis to small amplitude swings of the

pendulum. For large amplitude swings, the restoring force became non linear which led to a differential equation that is difficult to solve. As the amplitude increases, there is a lengthening of the period that is easy to measure but difficult to predict using calculus. However with the computer, it is as easy to use the exact force  $M_g \sin \theta$  as it is to use the approximate linear force  $M_g \theta$ . Thus with the computer you can predict the lengthening of the period and compare your results with experiment.

In Appendix I, we made a considerable effort to predict the effects of adding a time dependent driving force to a harmonic oscillator. The work paid off in that we got Equations 58 and 59 which provide a general description of resonance phenomena. With the computer you do not get these elegant formulas, but it is much easier to add a time dependent force and see what happens. In effect the computer solutions can be used as a laboratory to test the predictions of Equations 58 and 59. This provides an opportunity for a lot of project work.



#### Figure 33

Reproduction of Figure (10), showing an oscillating air cart. If the cart is displaced a distance x from equilibrium, there is a restoring force F = -kx. The force is measured by adding weights as shown.

## **English Program**

In Chapters 5 and 8 our usual approach for solving a new mechanics problem on the computer was to modify an old working program. But because the harmonic oscillator is an easy one dimensional problem, we will start over with a new program. Our general procedure has been to first write an English program that described the steps using familiar notation. Once we checked the steps to see that the program did what we wanted, we then translated the program into an actual computer language such as BASIC.

The English program for the oscillating air cart is shown in Figure 34. In the first section, we state the experimental constants, namely the mass M of the air cart and the spring restoring constant K. For this particular experiment, the cart has a mass M of 191 grams, and the spring constant K was 3947 dynes/cm. As indicated in Figure (33), the spring constant K was determined by tying a string to the mass, running the string over a pulley, and hanging weights on the other end. We got a linear force verses the distance curve like the one in Figure (9-4), and used the same method to find K.

In the next section of the program, we choose an explicit set of initial conditions. For this problem we start the cart from rest  $(V_0 = 0)$  at a distance 10 cm to the right of the equilibrium position  $(X_0 = 0)$ . The cart is released at time  $(T_0 = 0)$ .

In the lab we observed that the period of oscillation was about 1.5 seconds. Thus a calculational time step dt = .01 seconds gives us about 150 points for one oscillation, enough points for a smooth plot.

The calculational loop is similar to the one in the projectile motion program of Figure (5-18), page 5-16, except that for one dimensional motion we do not need vectors, and the old command

LET  $\vec{A} = \vec{g}$ 

is replaced by

LET F = -K\*XLET A = F/M On the next page we repeat this English program and show its translation into the computer language BASIC.

## **English Program**

! ----- Experimental constants LET M = 191 grams (cart mass) LET K = 3947 dynes/cm (spring constant)

! ----- Initial conditions

LET  $X_0 = 10 \text{ cm}$ LET  $V_0 = 0$  (release from rest) LET  $T_0 = 0$  (start clock)

! ----- Computer Time Step LET dt = .01

! ------ Calculational loop DO LET  $X_{new} = X_{old} + V_{old}*dt$ LET F = -K\*X (spring force) LET A = F/MLET  $V_{new} = V_{old} + A_{old}*dt$ LET  $T_{new} = T_{old} + dt$ PLOT X vs T LOOP UNTIL T > 15

END

#### Figure 34

English program for the motion of an oscillating cart on an air track.

## **English Program**

! ----- Experimental constants LET M = 191 grams (cart mass) LET K = 3947 dynes/cm (spring constant) ! ----- Initial conditions LET  $X_0 = 10 \text{ cm}$ LET  $V_0 = 0$  (release from rest) LET  $T_0 = 0$  (start clock) ! ----- Computer Time Step LET dt = .01! ----- Calculational loop DO LET  $X_{new} = X_{old} + V_{old} * dt$ LET F = -K\*X (spring force) LET A = F/MLET  $V_{new} = V_{old} + A_{old} * dt$ LET  $T_{new} = T_{old} + dt$ PLOT X vs T LOOP UNTIL T > 15

## END

## Figure 34 repeated

English program for the motion of an oscillating cart on an air track.

## The BASIC Program

Because no vectors are involved in the harmonic oscillator program, the translation into BASIC is almost automatic. Drop the subscripts "new" and "old", fix up the PLOT statement, add the plotting window commands, and you have the result shown in Figure (35a). Select **RUN** and you get the plot of oscillating motion shown in Figure (35b).

## **BASIC Program**

```
! ----- Plotting window
  SET WINDOW -5,15,-15,15
  PLOT LINES: -5,0; 15,0 ! T axis
  PLOT LINES: 0,-15; 0,15 ! X axis
! ----- Experimental constants
  LET M = 191
  LET K = 3947
! ----- Initial conditions
  LET X = 10
  LET U = 0
  LET T = 0
! ----- Computer Time Step
  LET dt = .01
! ----- Calculational loop
  DO
     LET X = X + U*dt
     LET F = -K*X
     LET A = F/M
     LET V = V + A*dt
     LET T = T + dt
     PLOT T,X;
  LOOP UNTIL T > 15
```

END

#### Figure 35a

BASIC program for the motion of an oscillating cart on an air track.



**Figure 35b** *Output of the BASIC program, showing the oscillation of the cart.* 

The plot of Figure (35b) nicely shows the sinusoidal oscillation, but does not tell us the numerical value of the period of oscillation. To determine the period, we modified the program as shown in Figure (36a). The main change is to replace the **PLOT** statement by a **PRINT** statement. To reduce the output, we included **MOD** statement (as described in Exercise 5-5, page 5-9) so that only every tenth calculated point would be printed. From the output shown in Figure (36b), we see that the period is close to 1.4 seconds. A more accurate value of the period can be obtained by not using the MOD statement and printing every value as shown in Figure (36c). From this section of data we see that the period is closer to 1.39 seconds.

#### Exercise 22

Show that the frequency of oscillation seen in the computer output of Figure (36) is consistent with the calculus derived equation

$\omega = \sqrt{K/IVI}$	ω	=	-√k/M
-------------------------	---	---	-------

Т	x	U	
.1	9.08409	-19.9673	
.2	6.32275	-35.8782	
.3	2.27695	-44.5005	
.4	-2.23141	-44.0825	
.5	-6.28646	-34.7092	
.6	-9.06442	-18.2848	
.7	-10.001	1.8542	
.8	-8.9058	21.6165	
.9	-6.00143	36.9875	
1.	-1.87788	44.8444	
1.1	2.62716	43.5913	
1.2	6.5985	33.4826	
1.3	9.22935	16.5719	
1.4	9.98527	-3.70531	
1.5	8.71269	-23.2298	
1.6	5.67013	-38.0352	
1.7	1.47568	-45.1137	
1.8	-3.01854	-43.0275	
1.9	-6.89955	-32.2002	
2.	-9.37893	-14.8315	
2.1	-9.95297	5.55026	
2.2	-8.50508	24.8045	
2.3	-5.32939	39.0196	
2.4	-1.07103	45.308	
2.5	3.4049	42.3921	
2.6	7.18913	30.8642	
2.7	9.5129	13.0663	
2.8	9.90412	-7.38598	
2.9	8.28333	-26.3378	
3.	4,97979	-39.9392	

#### Figure 36b

Numerical output.

! ----- Print labels PRINT "T", "X", "U" ! ----- Experimental constants LET II = 191 LET K = 3947 ! ------ Initial conditions LET X = 10LET U = 0LET T = 0! ----- Computer Time Step LET dt = .01 LET i = 0! ----- Calculational loop DO LET X = X + U\*dt LET F = -K\*XLET A = F/MLET V = V + A\*dt LET T = T + dt LET i = i+1IF MOD(i,10) = 0 THEN PRINT T,X,U LOOP UNTIL T > 3

END

## Figure 36a

Program for numerical output.

1.36	9.92701	4.55175
1.37	9.97253	2.49094
1.38	9.99744	.424976
1.39	10.0017	-1.64187
1.4	9.98527	-3.70531
1.41	9.94822	-5.7611
1.42	9.89061	-7.80499

#### Figure 36c

Detailed numerical output. By printing every calculated numerical value, we can more accurately determine the period of oscillation.

## **Damped Harmonic Motion**

In Figure (37a) we modified the projectile motion program of Figure (35a) to include damping The only change, shown in boxes in Figure (37a) is to replace

LET 
$$F = -K * X$$

by

LET F = -K\*X - b\*V

where we gave b the numerical value of 100 to get the result shown in Figure (35b).

#### **Exercise 23**

(This is more of an introduction to project work)

In our analysis of damped harmonic motion in Exercise 20, we predicted that the frequency for damped harmonic motion would be

$$\omega = \sqrt{\frac{k}{M} - \frac{b^2}{4m^2}}$$
 (51)

In the special case that

$$\frac{k}{M} = \frac{b^2}{4m^2};$$
  $b = \sqrt{4mk}$  (51a)

we get  $\omega = 0$  which is the case of critical damping, where oscillations cease.

Run the damped harmonic oscillator program of Figure (35a) for values of b near  $\sqrt{4mk}$  and show that oscillations cease when you get to this critical value.

Damped oscillation

1

```
! ----- Plotting window
SET WINDOW -5,15,-15,15
PLOT LINES: -5,0; 15,0 ! T axis
PLOT LINES: 0,-15; 0,15 ! X axis
```

! ------ Experimental constants LET M = 191 LET K = 3947 LET b = 100

! ----- Initial conditions LET X = 10 LET U = 0 LET T = 0

! ----- Computer Time Step LET dt = .01

! ------ Calculational loop D0 LET X = X + U\*dt LET F = -K\*X - b\*U LET A = F/M LET V = V + A\*dt LET T = T + dt PLOT T,X;

LOOP UNTIL T > 15

END

## Figure 37a

BASIC program for the damped harmonic motion.



**Figure 37b** *Plot of damped harmonic motion.* 

# Chapter 15 One Dimensional Wave Motion

In the last few chapters we have followed the straightforward procedure of identifying the forces on an object, setting the vector sum of the forces equal to the mass times acceleration, and solving the resulting equation. We began with problems like those involving circular motion where we knew the acceleration and could solve the equation immediately, the conical pendulum being an example. With oscillatory motion we ended up with a differential equation whose solution had to be guessed. The observation that oscillatory motion looks like circular motion viewed sideways helped greatly in this guess.

For damped and forced harmonic motion, it was not hard to write the differential equations, but the solutions involved mathematical functions and techniques that may not have been not familiar to the reader. In this chapter we are dealing with the subject of wave motion, where it turns out that the differential equation describing the motion has derivatives in both time and space. Setting up and solving such an equation requires mathematical discussions that are best left to a more advanced level course. Fortunately we can study the physics of wave motion without working with differential equations.

If we went through the effort to derive the differential equation for wave motion, we would end up with what is called a **wave equation**. Once you have a wave equation, you can guess a solution and plug in your guess just as we did for the simpler equations for oscillatory motion. For oscillatory motion, when we plugged in our guess  $\sin(\omega t)$ , we ended up with a simple equation  $\omega = \sqrt{k/m}$  for the frequency of the oscillation. For wave equations, if you plug in a guess representing a wave traveling through the medium, you end up with a simple equation for the speed of the wave.

There are some famous wave equations in physics. In 1860 James Clerk Maxwell combined the equations for electricity with those for magnetism and, to his surprise, ended up with a wave equation. He initially had no idea what the wave was, but he could calculate the speed of the wave. Whatever wave he was dealing with travelled at a speed of  $3 \times 10^8$  meters per second or 1 foot per nanosecond. As he knew of only one thing that travelled at that speed—light—he concluded that he had an equation for light waves and that the theory leading to this equation was the theory of light. He had discovered that light was an electric and magnetic phenomena.

In 1925, Louis De Broglie explained some baffling phenomena in atomic physics by proposing that electrons have a wave nature. Erwin Schrödinger then went further and derived a wave equation for the electron, an equation known as **Schrödinger's equation** that serves as the theoretical foundation for almost all of chemistry. In 1929 Paul Dirac constructed a relativistic generalization of Schrödinger's equation. The problem with Dirac's equation was that it had solutions for two different kinds of wave, one representing the electron and the other an unknown particle of the opposite charge. A particle similar to the electron but opposite in charge, the positron, was observed in a cloud chamber experiment carried out by Carl Anderson in 1933.

It turns out that the relativistic wave equation for all elementary particles has two solutions, one solution like the electron representing matter, the other, like the positron, representing **antimatter**. And the wave equations predict that if a matter particle encounters its corresponding antimatter particle, the two particles can annihilate each other. There is an entire world of antimatter, the existence of which was predicted by Dirac's wave equation.

With wave equations playing such an important role in physics, one might think it is unfortunate that we are not prepared to derive and solve wave equations. Actually that is a blessing. There are certain general, simple principles that apply to all forms of wave motion, principles that allow you to understand and predict many features of the behavior of waves. These principles apply not only to waves like water and sound waves whose behavior can be deduced from Newtonian mechanics, but to light and electron waves where Newtonian mechanics does not apply. Thus by learning these general principles of wave motion, you are developing a foundation in physics that goes beyond Newtonian mechanics.

The two basic principles of wave motion we will discuss in this text are the **principle of superposition** and the **Huygens principle**. The principle of superposition is a fancy way of saying that waves add. If two waves are moving through each other, they produce a total wave that is the sum of the two waves. Since waves can have negative amplitudes (troughs) this addition of waves can produce cancellation. Two waves running into each other can, under the right circumstances, cancel each other out. This cancellation is clearly a wave phenomena, particles are not expected to do that.

The other general principle of wave motion is Huygens principle, which tells us how waves spread out in space. In this and the next chapter we will focus our attention on one dimensional waves which do not spread out. Thus we do not need Huygens principle at this point. Later in Chapter 33, we discuss two and three dimensional waves and phenomena such as interference and diffraction. In that chapter we do everything using the principle of superposition and the Huygens principle. In that chapter no calculus is used and we obtain results that apply to a broad spectrum of phenomena, even to sub atomic particles where concepts like velocity and acceleration no longer have meaning.

## WAVE PULSES

We begin our discussion of wave motion with the wave pulses we described in Chapter 1 on Special Relativity. To create a wave pulse on a stretched rope, you flick the end of the rope and a pulse travels down the rope as shown in Figure (1-4) reproduced here. This is called a *transverse wave* because the particles in the rope move perpendicular or *transverse* to the direction of motion of the wave pulse.

With a stretched Slinky we were able to observe two different kinds of wave motion, the transverse wave seen in Figure (1-5) and a compressional wave seen in Figure (1-6). The compressional wave is also called a *longitudinal wave* because the particles in the spring



**Figure 1-4** *Wave traveling down a rope.* 



**Figure 1-5** *Transverse wave on a Slinky* 



**Figure 1-6** *Compressional wave on a Slinky.* 

are moving longitudinally or parallel to the direction of motion of the wave pulse.

Sound waves are usually compressional waves traveling through matter. A sound wave pulse in air can be viewed as a region of compressed gas where the molecules are closer together as shown in Figure (1). It is the region of compression that moves through the gas in much the same way as the region of compressed coils moves along the Slinky as seen in Figure (1-6).

To create the compressional wave on the Slinky we pulled back on the end of the Slinky and let go. This gives a small impulse directed down the Slinky. In much the same way we can use a loudspeaker cone to create the pressure pulse in the air column of Figure (1). Here the impulse can be provided by applying a voltage pulse to the speaker causing the speaker cone to suddenly jump forward. (If the speaker cone suddenly jumps back, you get a pulse consisting of a region of low pressure traveling down the tube.)

A transverse or sideways force in the medium tends to restore the medium to its original shape. For a transverse wave on a stretched rope, the tension on the rope provides the restoring force. For waves on the surface of a liquid, gravity or surface tension supplies the restoring force. But for waves passing through the bulk of a liquid or a gas, there are no transverse restoring forces and the only kind of waves we get are the compressional sound waves.



#### Figure 1

A sound wave pulse traveling down through a tube of air. The pulse consists of a region of compressed air where the air molecules are closer together. This region of compression moves through the gas much as the region of compressed coils moves along the Slinky in Figure 1-6. The main difference between a liquid and a solid is that in a liquid the molecules can slide past each other, while in a solid the molecules are held in place by molecular forces. These forces which prevent molecules from sliding past each other can also supply a transverse restoring force allowing a solid to transmit both transverse and compressional waves. An earthquake, for example, is a sudden disruption of the earth that produces both transverse waves called S waves and longitudinal or compressional waves called *P waves*. These waves can easily be detected using a device called a seismograph which monitors the vibration of the earth. It turns out that the S and P waves from an earthquake travel at different speeds, and will thus arrive at a seismometer at different times. By measuring the difference in arrival time and knowing the speed of the waves, you can determine how far away the earthquake was.

#### **Exercise 1**

The typical speed of a transverse S wave through the earth is about 4.5 kilometers per second, while the compressional P wave travels nearly twice as fast, about 8.0 kilometers per second. On your seismograph, you detect two sharp pulses indicating the occurrence of an earthquake. The first pulse is from the P wave, the second from the S wave. The pulses arrive three minutes apart. How far away did the earthquake occur?



#### Figure 2

Sketch of a simple seismograph for detecting earthquake waves. When the earth shakes, the mass tends to remain at rest, thus the pen records the relative motion of the stationary mass and shaking earth. (Building a seismograph is a favorite high school science fair project. Basically you suspend a large mass from springs and have a pen which is attached to the mass draw a line on moving stripchart paper as shown in Figure (2). When the earth shakes, the stripchart shakes with the earth, but the mass remains more or less stationary. The result is a squiggly line on the stripchart whose amplitude is the amplitude of vibration of the earth.

## **SPEED OF A WAVE PULSE**

One way to predict the speed of a wave is to set up the differential equation for the wave, plug in a traveling wave solution and let the equation tell you the speed. Without the wave equation we can in some cases deduce the speed of the wave using clever tricks. One example is the transverse wave on a rope, whose speed we will calculate now. Another is the speed of Maxwell's wave of electric and magnetic forces which we will discuss in Chapter 32.

To calculate the speed of a transverse wave on a rope, consider a wave pulse moving down a rope at velocity  $\vec{v}$  as shown in Figure (3a). To analyze the pulse, imagine that you are running along with the pulse at the same velocity  $\vec{v}$ . From your point of view, shown in Figure (3b), the pulse is at rest and the rope is moving back through the pulse at a speed v.

Now look at the top of the wave pulse. For any reasonably shaped pulse, the top of the pulse will be circular, fitting around a circle of radius r as shown in Figure (3c). This radius r is also called the *radius of curvature* of the rope at the top of the pulse.

Finally consider a short piece of rope of length  $\ell$  at the top of the pulse as shown in Figure (3d). If this piece of rope subtends an angle  $2\theta$  on the circle, as shown, then  $\ell = 2r\theta$  and the mass m of this section of rope is

$$m = \mu \ell = \mu 2 r \theta \qquad \frac{mass of short}{section of rope}$$
(1)

where  $\mu$  is the mass per unit length of the rope.

The net force on this piece of rope is caused by the tension T in the rope. As seen in Figure (3d), the ends

of the piece of rope point down at an angle  $\theta$ . Thus the tension at each end has a downward component T sin ( $\theta$ ) for a total downward force F<sub>y</sub> of magnitude

$$F_{y} = 2T \sin(\theta) \approx 2T\theta \qquad \begin{array}{c} downward \\ component of \\ tension force \end{array} (2)$$

If we keep the angle  $\theta$  small, just look at a very small section of the rope, then we can approximate sin ( $\theta$ ) by  $\theta$  as we did in Equation 2.



Figure 3a

Wave pulse, and an observer, moving to the right at a speed v.



#### **Figure 3b**

From the moving observer's point of view, the pulse is stationary and the rope is moving through the pulse at a speed v.



**Figure 3c** Assume that the top of the pulse fits over a circle of radius r.

The final step is to note that this section of rope is moving at a speed v around a circle of radius r. Thus we know its acceleration; it is accelerating downward, *toward the center of the circle*, with a magnitude  $v^2/r$ .

$$a_{y} = \frac{v^{2}}{r} \qquad \begin{array}{c} \textit{downward} \\ \textit{accelerationof} \\ \textit{sectionof rope} \end{array} \tag{3}$$

Applying Newton's second law to the downward component of the motion of the section of rope, we get using Equations 1, 2 and 3

$$F_{y} = ma_{y}$$

$$2T\theta = (\mu 2r\theta) \frac{v^{2}}{r}$$
(4)

Both the variables r and  $\theta$  cancel, and we are left with T =  $\mu v^2$ 

$$v = \sqrt{\frac{T}{\mu}}$$
speed of a wave pulse  
on a rope with tension T,  
mass per unit length  $\mu$ 
(5)

A result we stated back in Chapter 1.



The ends of the rope point down at an angle  $\theta$ , giving a net restoring force  $F_y = 2T \sin \theta$ .

## **DIMENSIONAL ANALYSIS**

In the above derivation of the speed of a transverse pulse on a rope, we avoided solving a differential equation by observing that the rope at the top of the pulse, from the moving observer's point of view, was moving with circular motion whose acceleration we know. For other kinds of wave pulses, particularly the compressional pulse seen in Figure (1-6), we do not have a simple circular motion, and a non calculus derivation of the wave speed becomes even more convoluted than the derivation we just went through. We could do it, but it is not worth the effort, especially since there are more straightforward ways of predicting wave speeds when one has the differential equation for the wave motion.

What we will do instead is use a technique called *dimensional analysis* to predict the speed of the wave. With dimensional analysis, you do not work out equations. Instead you determine what the relevant variables are, and then combine those variables in such a way that the dimensions are correct. If you have selected the correct variables, you get an answer that is correct to within a constant factor, and sometimes the correct answer.

To see how dimensional analysis works, let us first apply it to the example we just worked out-to find the speed of a transverse wave pulse on a rope. (For clarity, we will italicize the variable names. We will also use MKS units.) The first step is to do some experiments to find out what variables the speed depends upon. You choose a rope, stretch it, and soon discover that the speed of the pulse depends upon the tension T. Thus T is one of the variables. Then you try two ropes of the same length but different mass m, and discover that you get different wave speeds for the same tension. Thus the mass m is one of the relevant variables. Another experiment with 2 ropes of the same mass but different lengths, gives different wave speeds. Thus the rope length L is also important. Further experiments indicate that the speed of the pulse does not depend upon such variables as the color of the rope, the material from which it is constructed, or the time of day. Thus you conclude that the relevant variables and their dimensions are

$$T \frac{\text{kg m(meter)}}{\text{sec}^2}, \quad m \text{ kg}, \quad L \text{ m}$$
 (6)

From these variables we have to construct the velocity.

$$v \frac{\mathrm{m}}{\mathrm{sec}}$$
 (7)

The only variable with the dimensions of seconds in it is the tension T, thus T must be included in our formula for v. To get rid of kilograms, we must divide T by m to give

$$T \frac{\text{kg m}}{\text{sec}^2} * \frac{1}{m \text{ kg}} = \frac{T}{m} \frac{\text{m}}{\text{sec}^2}$$

We are getting there, but we must have the same power of meters and sec in order to get a velocity. If we multiply T/m by L meters, we get

$$\frac{T}{m} \frac{\mathrm{m}}{\mathrm{sec}^2} * L \,\mathrm{m} = \frac{TL}{m} \frac{\mathrm{m}^2}{\mathrm{sec}^2}$$

Finally we get the correct dimensions by taking the square root, giving

$$v = \sqrt{\frac{TL}{m}} \frac{\mathrm{m}}{\mathrm{sec}} = \sqrt{\frac{T}{\mu}} \frac{\mathrm{m}}{\mathrm{sec}}$$
 (8)

where we noted that  $\mu = m/L$  is the mass per unit length.

Equation 8 tells us that no matter what the theory is, if the only relevant variables are T, m and L, the speed of the wave must be proportional to  $\sqrt{T/\mu}$  for the dimensions to work out. We may have missed a factor of 1/2 or  $2\pi$ , but the functional dependence must be right.

Let us now use dimensional analysis to predict the speed of the compressional Slinky pulse shown in Figure (1-6), or any compressional pulse on a stretched spring. Since a stretched spring has a tension *T* and a mass per unit length  $\mu$ , one might guess that  $\sqrt{T/\mu}$  could also be the formula for the compressional wave. However compressional and transverse waves do not have the same speed. Even more important, you can get different wave speeds for the same value of  $T/\mu$ , by using different springs. It turns out that the tension *T* is not a relevant variable.

Compressional waves depend upon the stiffness of a material, not the tension. For example a compressional sound pulse will travel down a steel rod whether or not the rod is under tension. Pulling on the ends of the steel rod does not noticeably change the speed of the sound pulse. Increasing the tension in a spring stretches the spring and therefore changes the mass per unit length. It is the change in mass per unit length, not the change in tension, which affects the speed of the compressional pulse.

What variable is related to the inherent stiffness of a spring? The one that comes to mind is the spring constant k that appears in Hooke's law

$$F = -kx;$$

$$|k| = \frac{F}{x} \frac{\text{newtons}}{\text{meter}} \qquad Hooke's \, law$$
(9)

The stiffer the spring, the greater the spring constant *k*.

Suppose we decide, after enough experimentation, that the relevant variables for the compressional pulse on a spring are the spring constant k, spring mass m and spring length L. We obtain the dimensions of k from Hooke's law,

$$k \frac{\text{newtons}}{\text{meter}} = k \frac{\text{kg} \cdot \text{m/sec}^2}{\text{m}} = k \frac{\text{kg}}{\text{sec}^2}$$

thus we have to construct a quantity with dimensions m/sec from the variables

$$k \frac{\mathrm{kg}}{\mathrm{sec}^2}$$
,  $M \mathrm{kg}$ ,  $L \mathrm{m}$ 

The only way we can do it is to divide k by m to get rid of kilograms and multiply by  $L^2$  to get

$$\frac{kL^2}{m} \frac{\mathrm{m}^2}{\mathrm{sec}^2}$$

Taking the square root gives a quantity with the dimensions of a velocity

$$v = \sqrt{\frac{kL^2}{m}} = \sqrt{\frac{kL}{\mu}} \quad speed of a compressional wave on a spring (10)$$

where again  $\mu = m/L$  is the mass per unit length. This is our prediction for the speed of a compressional wave on a spring. The actual speed could differ by a constant factor like 2, but it must have this functional dependence if we are correct in our assumption that the only relevant variables are k, m and L.

In the formula  $v = \sqrt{kL/\mu}$ , the appearance of the product *kL*, rather than *k* alone may at first seem surprising. But it turns out that the inherent stiffness of

a spring is proportional to kL and not just k, with the result that the speed of the pulse is related to the stiffness, as we suspected.

To see why the inherent stiffness is related to kL, imagine that we wind a long spring and cut it in half to create two identical springs of length  $L_1$ . As shown in Figure (4a), if we apply a force F to one of the springs, and measure the distance  $\Delta x$  that the spring stretches, we can use Hooke's law to calculate that the spring constant  $k_1$  is given by

$$k_I = \frac{F}{\Delta x} \tag{11}$$

Now attach the two springs back together and stretch the combination with the same force *F* as shown in Figure (4b). Since each spring feels the same force *F*, each stretches a distance  $\Delta x$ , and the pair stretch a distance  $2\Delta x$ . Thus from Hooke's law the  $k_2$  of the combination is given by

$$k_2 = \frac{F}{2\Delta x} = \frac{1}{2} \left( \frac{F}{\Delta x} \right) = \frac{k_I}{2}$$
(12)

where we used  $k_1 = F/\Delta x$  from Equation 11.

$$\begin{array}{c} \longleftarrow L_{1} \longrightarrow \\ \text{(unstretched)} & \text{(where the equation of the equation of$$

#### Figure 4a

Measuring the spring constant of a spring.

(unstretched) 
$$4L_1 \longrightarrow 2L_1 \longrightarrow 2L_1 \longrightarrow 2L_1 \longrightarrow 2L_2 \longrightarrow$$

Figure 4b

Measuring the spring constant of two connected springs.

When we attach two identical springs together, we end up with a longer spring but we are not changing the inherent stiffness of the spring. More importantly we do not change the speed of the wave pulse. Connecting the two springs and keeping the tension F the same merely gives the pulse a longer distance to travel.

Note, however, that when we attach the two springs, the spring constant is cut in half, but the length is doubled, with the result that the product kL is unchanged. Explicitly we have

$$k_2 L_2 = \left(\frac{k_1}{2}\right) \left(2L_1\right) = k_1 L_1 \tag{13}$$

It should now appear more reasonable that the speed of the wave pulse should be given by  $\sqrt{kL/\mu}$ . Both the quantity kL, and the mass per unit length  $\mu$  are inherent properties of the spring that do not depend on the length of the spring. It is thus reasonable that the speed of the wave pulse should also involve only these variables.

#### **Project Suggestion**

We have spent some time discussing the speed of pulses on a stretched spring for two reasons. One is that we used these pulses as our main example of wave motion in our introduction to special relativity in Chapter 1. The second is that measuring the speed of pulses on a spring makes a nice project, not much equipment is needed, and you can fairly easily measure the variables needed to test Equation 10.

An additional advantage is that Equation 10 might or might not be right. Since it was derived by dimensional analysis, it could be off by a constant factor like 1/2 or  $2\pi$ . Therefore you have the challenge of determining whether or not there are some missing constant factors.

We expect that the wave speed should be proportional to  $\sqrt{kL/\mu}$ , and if this does not turn out to be correct, we have made some mistake in our analysis of what variables are important. For example, in our analysis, we said nothing about the unstretched length  $L_0$  of the spring. Should  $L_0$  also appear in the formula for  $v_{wave}$ ? The way to find out is to do some experiments.

The experiments are made a bit easier by noting that

$$v = \sqrt{\frac{kL}{\mu}} = \sqrt{\frac{kL^2}{m}} = L\sqrt{\frac{k}{m}}$$

where we used  $\mu = m/L$ .

## SPEED OF SOUND WAVES

The quantity kL that appeared in our formula for the speed of a wave pulse is essentially the stiffness of a unit length spring. By stiffness we mean the ratio of the force applied to stretch a unit length of spring, to the amount of stretch  $\Delta x$  that we get. The same ideas also apply to stretching a steel rod or any one dimensional object that has an elasticity and obeys Hooke's law.

In engineering texts, the force applied to a unit length, area, or volume is given the generic name *stress*, and the resulting displacement that the stress causes is called a *strain*. The ratio of the *stress to the strain*, is called the *modulus*. For our spring, the stress is the tension force *F*, and the strain is the change in length per unit length, or  $\Delta x/L$ . The ratio of the stress to the strain,  $F/(\Delta x/L) = FL/\Delta x$  is called **Young's modulus**. From Hooke's law,  $F/\Delta x = k$ , thus Young's modulus is  $FL/\Delta x = kL$ , the quantity we have been discussing.

When we have a compressional wave in a gas, we can think of the compression as being caused by a pressure pulse that travels through the gas. In the region where the gas is compressed, there is a slight excess pressure. The speed of the wave pulse depends upon the response of the gas to this excess pressure.

Using the engineering terminology, the excess pressure  $\Delta P$  represents the stress and the fractional change in volume,  $\Delta V/V$  the corresponding strain. In this case the ratio of the stress  $\Delta P$  to the strain  $\Delta V/V$  is called the **bulk modulus B**. The formula for B is thus

$$B = \Delta P / (\Delta V / V) \quad bulk \ modulus \tag{15}$$

In Chapter 17 we will discuss the concept of a pressure in a gas, and see how changes in pressure are related to changes in volume. Until we get to that chapter, any detailed discussion of the concept of bulk modulus is premature. What we will do now is assume that it is the bulk modulus *B* essentially represents the "stiffness" of the gas and should appear in the formula for the speed of a sound wave. We will then use dimensional analysis to figure out what the formula should be. In the ratio  $B = \Delta P / (\Delta V / V)$ , the denominator  $\Delta V / V$  is dimensionless, thus *B* has the dimensions of pressure which is a force per unit area.

$$B \frac{\text{newtons}}{\text{meter}^2} = B \frac{\text{kg m}}{\sec^2 \text{m}^2} = B \frac{\text{kg}}{\sec^2 \text{m}} \quad (15)$$

To construct a quantity involving *B* that has the dimensions of a velocity, we have to get rid of the kilograms by dividing by some quantity related to the mass of the gas. The only reasonable choice is the gas density  $\rho \text{ kg/m}^3$ , thus we now have

$$\frac{B \text{ kg/(sec}^2 \text{ m})}{\rho \text{ kg/m}^3} = \frac{B}{\rho} \frac{\text{m}^2}{\text{sec}^2}$$
(16)

which is the square of a velocity. Taking the square root, we get

$$v = \sqrt{\frac{B}{\rho}} \qquad speed of a \\ sound wave \qquad (17)$$

a result we stated back in Chapter 1.

Equation 17 holds not only for a gas, but also for compressional sound waves in a liquid and a solid. Liquids and solids, being far more incompressible than a gas, have a much greater bulk modulus B and therefore higher speeds of sound. The speed of sound in various substances is given in Table 15-1.

Substance	Speed of Sound	
	in meters/sec	
Gases (at atmos	pheric pressure)	
Air at 0° C	331	
Air at 20° C	343	
Helium at 20° C	965	
Hydrogen at 20° C	1284	
Lie	quids	
Water at 0° C	1402	
Water at 20° C	1482	
Sea water at 20° C	1522	
Se	olids	
Aluminum	6420	
Steel	5941	
Granite	6000	
Nuclear matter	near c	
Table 15-1		

From this table we see that the speed of sound is considerably higher in the light gasses helium and hydrogen than in the more dense gas air. (The difference in density is why helium and hydrogen filled balloons float in air.) The compressibility or bulk modulus is usually the same for all gases at a given pressure, thus the higher speed in hydrogen or helium is due to the lower density, as we would expect from the formula  $v = \sqrt{B/\rho}$ .

#### Exercise 2

Steel is much stiffer than aluminum. (You make much better springs from steel than aluminum.) Yet the speed of sound is greater in aluminum than steel. Why?

#### Exercise 3

You tap the end of a 10 meter long steel rod with a hammer. How long before the tap can be detected at the other end of the rod?

#### **Exercise 4**

## A dimensional analysis problem that you should attempt now.

When working with the theory of electric and magnetic phenomena, using the MKS system of units, one encounters two rather mysterious constants labeled  $\varepsilon_0$  (epsilon naught) and  $\mu_0$  (mu naught). The constant  $\varepsilon_0$  appears in the formula for electric forces, and  $\mu_0$  in the formula for magnetic forces. These constants have the following dimensions

$$\epsilon_0 \frac{(\text{coulomb})^2 (\text{seconds})^2}{\text{kilogram}(\text{meter})^3}$$
 (18)

$$\mu_0 \frac{\text{kilogram meter}}{(\text{coulomb})^2}$$
(19)

where a coulomb is a unit of electrical charge. The numerical values of  $\varepsilon_0$  and  $\mu_0$  are to be found on the inside cover of this text along with other important physical constants.

(a) what combination of the constants  $\epsilon_0$  and  $\mu_0$  have the dimensions of a velocity?

(b) from the numerical value of this velocity, what do you think it is the velocity of?

## LINEAR AND NONLINEAR WAVE MOTION

Few sights are more awesome than the crashing of ocean rollers on a rocky beach during a storm. The waves seen in Figure (3) of Chapter 1, produced by hurricane Bertha hundreds of miles out to sea, were crashing against the rocky shores of Mt. Desert Island, Maine, in July 1990. Hundreds of tourists and local television station reporters were at the beach to observe the event. At one spot, called 'Thunder Hole', the crashing waves created a loud boom and a geyser of water that went 40 or 50 feet in the air.

A very different sight are the circular ripples emerging from where raindrops have hit a puddle of water, seen in Figure (1-2), reproduced here. The special feature of these ripples is that they maintain their identity as they move through each other. They are still circular waves even after moving through other waves.



**Figure 1-3** *This ocean wave from Hurricane Bertha (July 31, 1990).* 



**Figure 1-2** *Rain drops creating circular waves on a puddle.* 

Conceptually we can separate all wave motion into two classes. There are the relatively smooth waves that can pass through each other like the circular ripples of Figure (1-2), and the relatively wild waves that crest, crash and change their shapes. The relatively smooth waves all obey what is called a *linear* wave equation. The properties of these waves are well understood and their behavior easy to predict. The wild waves obey *nonlinear* wave equations. We know very little about the behavior of nonlinear waves, and in most cases find it very difficult to make predictions about their behavior. (Ocean waves, for example, are linear until they start to crest. When you see whitecaps, the waves have become nonlinear.)

In this text we will restrict our discussion to the smooth, linear waves that behave like the circular ripples. Fortunately, most kinds of wave motion we encounter in nature, including almost all examples of light waves and the probability waves of quantum mechanics, are linear and therefore relatively easy to analyze. But there are growing applications for nonlinear waves, particularly in the field of laser optics.



**Figure 5** *Two wave crests running into each other add up to produce a bigger crest.* 

## THE PRINCIPLE OF SUPERPOSITION

Figure (1-2) illustrates one aspect of linear wave motion. The waves can move through each other and emerge undisturbed. The waves are still circular and unbroken after they have crossed.

There is another simple feature of this wave motion that is a bit harder to see from that picture. While the waves are crossing, they produce a wave whose height is the sum of the heights of the individual waves. If two crests are moving through each other, the crests add to produce a higher crest. Two troughs produce a deeper trough, and a crest and a trough will tend to cancel as they move through each other.

This adding of the heights of crossing waves is more easily illustrated for the case of one dimensional wave pulses traveling down a rope. In Figure (5), two similar crests add together to produce a doubly high crest for an instant. In Figure (6) we see that a similar shaped crest and trough will cancel at the instant they are together. Figure (7) illustrates the idea that as any two wave shapes move through each other, they produce a wave shape whose height at any point along the rope is the sum of the heights of the individual waves moving through each other. The concept that waves can maintain their identity as they move through each other, and that they produce a resultant wave whose height or *amplitude* is the sum of the heights or amplitudes of the individual waves, this concept is known as the *principle of superposition*. In more colloquial language, the principle of superposition says that *waves add*. The principle of superposition is one of the key concepts of linear wave motion. It distinguishes linear from nonlinear wave motion. When nonlinear waves interact, you get something different than the simple sum of the two waves.

Before leaving our discussion of the principle of superposition, we wish to take one further look at part (c) of Figure (6). That is the point where an equal shaped crest and trough are right on top of each other, precisely cancelling each other out. This kind of cancellation of waves is a common feature of wave motion. In fact, it is what distinguishes *wave motion* from what we have been calling *particle motion*. If two particles run into each other, they do not cancel like the waves of Figure (6). They bounce or crash but not cancel.





**Figure 6** When a crest meets a trough, there is a short time when the waves cancel.



## SINUSOIDAL WAVES

If you ask someone to describe wave motion, they are likely to picture water waves and sketch a curve that looks like a sine wave. A sine wave represents just one of many possible shapes for a wave. But it is an important shape because it is often seen in nature and it is easy to handle mathematically. We will see shortly that any arbitrary wave shape can be constructed from sine waves, thus the sine wave can be thought of as a basic building block of wave motion.

To relate the mathematical sine function to wave motion, recall our definition of sine function shown back in Figure (14-4). The point of that figure is that the sine function is the sideways projection of circular motion. As the arrow rotates at an angular velocity  $\omega$ , the angle  $\theta$  that the arrow has rotated increases as  $\theta = \omega t$ . On the right we have graphed the height of the rotating arrow as a function of the angle  $\theta = \omega t$  to obtain a sine curve.

To actually create the sine wave shape seen in Figure (14-4), you can start shaking one end of a long rope as shown in Figure (8). If you move your hand up and

down with a sinusoidal oscillation, a sinusoidal shaped wave will start traveling down the rope, at a speed  $v_{wave} = \sqrt{T/\mu}$ . This creates an example of what is called a *traveling sine wave*.

The problem with creating traveling sine waves on a rope, is that the wave reaches the end of the rope, reflects, and moves back through the incoming wave, complicating the situation. A better example of traveling sine waves can be seen on the surface of a lake or the ocean where there is plenty of room for the waves to move before they strike an object or a shore.

There are two distinct ways to view a traveling sine wave. One is to move along with the wave. Then all you see is a stationary sinusoidal shape. The other is to stand still and let the wave pass by you. Then you will see the wave oscillate up and down as successive crests and troughs pass by you. This is illustrated in Figure (9) where we have sketched a traveling sinusoidal water wave passing a fixed post in the water. If you move along with the wave, then the shape of the wave does not change. But if you look at the post, the level of the water is moving up and down with a sinusoidal oscillation.



## Wavelength, Period, and Frequency

We usually describe a wave in terms of its wavelength  $\lambda$ , frequency f or period T. The easy way to remember how to go back and forth between these quantities, is to use dimensions.

When we view the shape of a traveling sine wave, the predominant feature is the wavelength, the distance  $\lambda$  between crests shown in Figure (8). Considering that one full cycle of the wave fits between the crests, we can assign the dimensions of *meters per cycle* to  $\lambda$ .

$$\lambda \frac{\text{meters}}{\text{cycle}}$$
 wavelength

When we let the wave pass by us and view the up and down motion of the surface, we see an oscillation whose period is T *seconds per cycle* and frequency is f *cycles per second*.

Since the period T is the length of time it takes one wavelength  $\lambda$  of the wave to pass by at a speed  $v_{wave}$ , we have (*distance = speed times time*)

$$\lambda \frac{\text{meters}}{\text{cycle}} = v_{\text{wave}} \left( \frac{\text{meters}}{\text{second}} \right) T \left( \frac{\text{second}}{\text{cycle}} \right)$$
 (20)

By assigning the dimensions meter/cycle to  $\lambda$  and sec/cycle to T, we can get the relationship  $\lambda = v_{wave}T$  from dimensions without having to memorize formulas, or even having to think very much.

#### Figure 9

Traveling sine wave on the surface of water. The sine wave shape moves as a unit along the surface at a speed  $v_{wave}$ . But if we look at a fixed post in the water, the water level at the post oscillates up and down with a sinusoidal oscillation. As an example of using dimensions to derive a formula, let us see if we can get a formula for the frequency f of a wave of wavelength  $\lambda$ . The idea of using dimensions is to try something, then see if the dimensions match. If they don't match, change the formula until they do. As a guess, let us try the formula

 $f = v_{wave}\lambda$  guess

Putting in dimensions, we have

,

$$f\left(\frac{\text{cycles}}{\text{sec}}\right) = v_{\text{wave}}\left(\frac{\text{meters}}{\text{sec}}\right) \times \lambda\left(\frac{\text{meters}}{\text{cycle}}\right)$$
$$= v_{\text{wave}}\lambda\frac{\text{meters}^2}{\text{sec cycle}}$$

Clearly the dimensions do not match. We have to change the formula so that meters cancel and we get cycles upstairs. This can be done if we move  $\lambda$  downstairs, giving

$$f\left(\frac{\text{cycles}}{\text{sec}}\right) = \frac{v_{\text{wave}}(\text{meters/sec})}{\lambda(\text{meters/cycle})}$$
$$= \frac{v_{\text{wave}}}{\lambda}\left(\frac{\text{cycles}}{\text{sec}}\right)$$
(21)

which works. Thus the correct formula is  $f = v_{wave}/\lambda$ , a result that can be a bit tricky to figure out other ways.



#### Angular Frequency ω

In Figure (14-4), we reminded ourselves that a sinusoidal oscillation is equivalent to the sideways view of circular motion. If the vector on the left side of Figure (14-4) is rotating with an angular velocity  $\omega$  *radians per cycle*, we get one rotation, one period T of the oscillation, when the angle  $\theta = \omega t$  goes from 0 at t = 0, to  $2\pi$  at t = T. Thus at the end of one period,  $\theta = \omega T = 2\pi$ .

Again we can avoid memorizing new formulas by using dimensions. We note that  $2\pi$  is the number of radians in a full circle or cycle. Thus we will assign to it the dimensions

$$2\pi \frac{\text{radians}}{\text{cycle}}$$
 (23)

Then to find the formula, for example, for the wave's angular frequency  $\omega$  in terms of the wave's period T, we have

$$\omega\left(\frac{\text{radians}}{\text{sec}}\right) = \frac{2\pi\left(\text{radians / cycle}\right)}{T\left(\text{sec / cycle}\right)}$$
$$= \frac{2\pi\left(\frac{\text{radians}}{\text{sec}}\right)$$
(23)

#### Exercise 5 (Do this one now.)

For a traveling sine wave moving at a speed  $\nu_{wave}$  , use dimensions to find

- (a)  $\lambda$  in terms of  $v_{wave}$  and  $\omega$
- (b)  $v_{wave}$  in terms of  $\lambda$  and T
- (c) T in terms of  $\lambda$  and  $\omega$
- (d) f in terms of  $\omega$

#### **Spacial Frequency k**

When we let a traveling wave pass by us, we observe a sinusoidal oscillation in *time*. This oscillation can be described in terms of the number of seconds in each cycle (T seconds/cycle), in terms of the frequency (f cycles/second) or the angular frequency ( $\omega$  radians/second).

If instead we look at the whole wave at one instant of time, or move along with the wave, we see a sinusoidal oscillation in *space*. We have described this spacial oscillation in terms of the wavelength, the number of meters in each cycle ( $\lambda$  meters/cycle). What we are missing is a spacial analogy to frequency, the *number of cycles or radians per meter*.

By dimensions we immediately see that

$$\frac{1}{\lambda \frac{\text{meters}}{\text{cycle}}} = \frac{1}{\lambda} \frac{\text{cycles}}{\text{meter}}$$

is the special analogy to the time frequency f, and that

$$\frac{2\pi (radians/cycle)}{\lambda (meters/cycle)} = \frac{2\pi}{\lambda} (\frac{radians}{meter})$$

is the spacial analogy to the angular frequency  $\omega$ .

In physics texts, it is not common to use a special symbol to designate the spacial frequency  $1/\lambda$  (cycles/meter), but it is standard practice to designate the angular spacial frequency  $2\pi/\lambda$  (radians/meter) by the letter k

$$k\frac{\text{radians}}{\text{meter}} \equiv \frac{2\pi \text{ radians}}{\lambda \text{ meter}} \quad \substack{\text{spacial} \\ \text{frequency}}$$
(24)

The standard name for k is the lackluster expression *wave number*, which says very little about the quantity. Instead we will refer to k as the *spacial frequency* of the wave. The higher the spacial frequency k, the more radians we get in a meter, just as the higher the time frequency  $\omega$ , the more radians we get in one second.

When you first study wave motion, it may be an irritating complication to have two kinds of frequency, f cycles/sec and  $\omega$  radians/sec (or two spacial frequencies  $1/\lambda$  cycles/meter and k radians/meter). Why not stick with cycles which are much easier to visualize than radians? The answer is that in the formulas for sine waves, the sine function basically has to be expressed in terms of an angle, as in sin  $\theta$ , and radians are an angle. To convert time t to an angle, we multiply by  $\omega$  as in

$$\theta$$
 radians =  $\omega \frac{\text{radians}}{\text{sec}} \times t$  sec  
=  $\omega t$  radians (25a)

while to convert the distance x to an angle we multiply by k as in

$$\theta$$
 radians =  $k \frac{radians}{meter} \times x$  meters  
= kx radians (25b)

Using Equations 25a or 25b, we can express the single function  $\sin \theta$  either as  $\sin \omega t$ , a sine wave in time shown in Figure (10a), or as  $\sin kx$ , a sine wave in space shown in Figure (10b). From these graphs we can see that when  $\omega t$  gets up to  $2\pi$ , we have one period T, and when kx gets up to  $2\pi$  we have one wavelength  $\lambda$ .

#### Exercise 6 (Try this now.)

You have a traveling sine wave moving at a speed  $v_{wave}$ . Using dimensions find the formula for  $v_{wave}$  in terms of the wave's time frequency  $\omega$  and spacial frequency k.



#### Figure 10

Sine waves in time and space.

## **Traveling Wave Formula**

Thus far we have formulas for a time varying sine wave  $\sin \omega t$ , and a space varying wave  $\sin kx$ . Now we want a formula for a traveling sine wave whose amplitude varies in both space and time. The answer turns out to be

$$y = \sin \theta = \sin(kx - \omega t) \frac{traveling}{sine wave}$$
 (26)

What we will do is show that this formula represents a sine wave moving down the x axis.

Figure (11) shows a sinusoidal shape that is moving down the x axis at a speed  $v_{wave}$ . If we describe the wave by the function  $\sin \theta$ , then it is the origin  $\sin \theta = 0$  that moves down the axis at a speed  $v_{wave}$ . Thus what we need is a formula for  $\theta$  so that when we set  $\theta = 0$ , that point does move down the x axis at the desired speed.

The answer we gave in Equation 26 suggests that the correct formula for  $\theta$  is

$$\theta = \mathbf{k}\mathbf{x} - \boldsymbol{\omega}\mathbf{t} \tag{27}$$

Setting  $\theta = 0$  we get

$$\theta = 0 = kx - \omega t$$
;  $kx = \omega t$   
 $x = \left(\frac{\omega}{k}\right) t$  (28)

But if the  $\theta = 0$  point travels at a speed  $v_{wave}$ , then after a time t, it has traveled a distance x given by

$$\mathbf{x} = \mathbf{v}_{wave} \mathbf{t} \tag{29}$$

Comparing Equations (28) and (29), we see that the point  $\theta = 0$  moves along the x axis at a speed

$$v_{wave} = \frac{\omega}{k}$$
(30)

If you did Exercise 6, you recognize that the quantity  $\omega/k$  has the dimensions of a velocity

$$\frac{\omega}{k} \frac{\text{radian/sec}}{\text{radian/meter}} = \frac{\omega}{k} \frac{\text{meter}}{\text{sec}} = v_{\text{wave}}$$
(31)

Thus the origin does move down the x axis at a speed  $v_{wave}$ , and the formula  $(kx - \omega t)$ , is our desired traveling wave formula.

#### **Exercise 7**

Explain what kind of a wave is represented by the formula  $y = sin(kx + \omega t)$ .



**Figure 11** The cycle begins at  $\theta = 0$ .

## **Phase and Amplitude**

The Equation 26 for a traveling wave can be generalized by noting that the wave can have an arbitrary amplitude A, and an arbitrary constant phase angle  $\phi$ to give

$$y = A \sin(kx - \omega t + \phi)$$
 (32)

The amplitude A just makes the sine wave bigger or smaller, and the phase angle  $\phi$  shifts the sine wave to the left or right.

To see precisely how the phase angle  $\phi$  shifts the sine wave, we have in Figure (12) compared  $\sin \theta$  and  $\sin (\theta + \phi)$ . The function  $\sin (\theta + \phi)$  crosses zero when the angle  $(\theta + \phi) = 0$  or at  $\theta = -\phi$ . Thus adding a phase angle  $\phi$  shifts the sine wave back a distance  $-\phi$  radians. If, for example, we set  $\phi = \pi/2$ , the wave is shifted back 1/4 of a wavelength, and we have converted a sine wave into a cosine wave.



**Figure 12** Adding a phase angle  $\phi$  shifts the wave back a distance  $\phi$ .

## **STANDING WAVES**

In addition to the waves traveling down a rope, another kind of wave pattern that is easy to achieve are those shown in Figure (13). All you have to do is shake the end of the rope at the right frequency and one of these waves will appear. Change the frequency and you can change to one of the other patterns.

The waves in Figure (13) are called *standing waves* because the pattern does not move along the rope. The points of zero amplitude, the points called *nodes* of the wave, stay at fixed positions while the rope between nodes oscillates back and forth.

The difference between the wave patterns is characterized by the number of nodes. In Figure (13) all the waves have nodes at the ends, and there are zero, one, two and three nodes in between as we go from the left to the right pattern. The two kinds of waves on a rope, the traveling wave of Figures (8) and (9) and the standing wave of Figure (13) are closely related to each other. A careful demonstration shows how traveling waves can turn into standing waves.

If you start shaking a rope a traveling wave starts down the rope as shown in Figure (8). After a while the wave reaches the other end of the rope, is reflected, and starts moving back the other way. This reflection is most easily seen if you send a single pulse down the rope so that you can see it bounce off the fixed end and come back to you.

If you send a series of pulses down the rope, if you create a traveling sine wave, then the reflected pulses have to move back through the pulses that are still coming in. You now have the superposition of two traveling waves moving through each other in opposite directions.



**Figure 13** *Standing waves on a rope.* 

You might think that the sum of two traveling waves moving through each other could lead to a complex pattern, and when you try it in a demonstration it often looks that way. The problem with a demonstration is that the reflected wave reaching your hand can reflect again and you begin to build up a mixture of many waves.

If you do the demonstration carefully, however, you can observe a simple result. *The sum of the traveling wave and the reflected wave, moving through each other, is a standing wave.* The addition of two traveling waves of equal amplitude and wavelength moving in opposite directions through each other is illustrated in Figure (14). The two traveling waves are shown on lines (a) and (b) at five different times t = 0, T/4, T/2, 3T/4 and T, where T is the period of the waves. On the bottom line (c) we have added the amplitudes of the two traveling waves to get a picture of the amplitude of the resulting wave.

In the first frame t = 0, the two waves match exactly, producing a sum that has twice the amplitude of either traveling wave. A quarter of a period later, at t = T/4, the traveling waves are precisely opposite each other. And the sum is zero all along the wave. This never happens in a traveling wave. A traveling wave is never completely flat, there is always a crest moving along.

At time t = T/2, half a period later, the traveling waves again line up producing a wave of twice the amplitude. Note now that the points in the sum wave that were below the axis at t = 0 are now above the axis at t = T/2, and vice versa. At time t = 3T/4 the traveling waves are again out of phase and add up to zero. At t = T, we are back to where we started.

From line (c) of Figure (14), we see that the nodes of the sum wave remain stationary and the rope between the nodes oscillates up and down. This is exactly what we see in the photographs of the standing rope waves in Figure (13).



#### Figure 14

Making a standing wave out of two traveling waves on a rope. In the top line (a) we show a traveling wave moving to the right. A section of rope is shown at times t = 0, T/4, T/2, 3T/4 and T, where T is the period of the wave. In (b) we have, in the same section of rope, five views of a traveling wave moving to the left. In (c), we have added the two traveling waves and get a standing wave with stationary nodes. To demonstrate the addition, at the center of each time frame we have drawn arrows to show the height of the wave at that point. The length of the bottom arrow is the sum of the lengths of the upper two arrows. To add two waves, you add up the heights at each point along the wave.

Using a trigonometric identity, we can show mathematically that the sum of two traveling waves moving through each other creates a standing wave. The formula for a sine wave traveling to the right is from Equation 26.

$$y_{\text{moving right}} = A \sin(kx - \omega t)$$
 (26)

If you worked Exercise 6, you found that the formula for a similar wave moving left is

$$y_{\text{moving left}} = A \sin(kx + \omega t)$$
 (33)

The formula for the sum of the two waves is

$$y_{sum wave} = A \sin(kx - \omega t) + A \sin(kx + \omega t)$$

The trigonometric identity we will use is

$$\sin(a \pm b) = \sin(a)\cos(b) \pm \cos(a)\sin(b) \quad (34)$$

This gives

$$\sin (kx - \omega t) = \sin (kx) \cos (\omega t) - \cos (kx) \sin (\omega t)$$
$$\sin (kx + \omega t) = \sin (kx) \cos (\omega t) + \cos (kx) \sin (\omega t)$$

Add these two sine waves, the  $\cos(kx)\sin(\omega t)$  terms cancel and we are left with

$$y_{sum wave} = 2A \sin kx \cos \omega t$$
 (35)

To interpret Equation 35, write it in the form

$$y_{sum wave} = A(x) \cos \omega t$$
 (36a)

where the x dependent amplitude A(x) is

$$A(x) = 2A \sin kx \tag{36b}$$

Equation 36a tells us that the entire wave is oscillating in time as  $\cos(\omega t)$ . However the amplitude of the oscillation depends upon the position x long the wave. Equation (36b) tells us how the amplitude varies with position. It varies sinusoidally as sin kx, with nodes permanently located at the points where sin kx =0. This sinusoidal variation in the amplitude along the wave is clearly seen in the photographs of the standing waves on the rope, Figure (13).

## WAVES ON A GUITAR STRING

Perhaps the clearest example of standing waves are the waves on the strings of a stringed instrument such as the guitar. The advantage of working with these waves is that you get to both see the shape of the wave and hear its frequency.

The shape of guitar string waves are the same as the standing rope waves in Figure (13). In Figure (15), we have sketched the allowed standing wave patterns on a string of length L. Because the string is fixed at the ends, we can only have waves with nodes at the ends.



#### Figure 15

Allowed standing waves on a guitar string. The formula for the wavelength of the nth harmonic is seen to be  $\lambda_n = 2L/n$ .

The wave with no nodes between the ends is called the *fundamental* or *first harmonic*. Its wavelength is 2L, twice the length of the string. In the second harmonic, with one node in the middle, a full wave fits on the string at one time and we have  $\lambda_2 = L$ . Each time we add a node we go up to one higher harmonic. (*The second harmonic is also called the first overtone, etc.*)

The general formula for the wavelength of the nth harmonic can be seen is we write the progression of wavelengths in the form

$$\lambda_1 = \frac{2L}{1}; \ \lambda_2 = \frac{2L}{2}; \ \lambda_3 = \frac{2L}{3}, \ \text{etc.}$$

It is clear that the formula for  $\lambda_n$  is simply

$$\lambda_{n} = \frac{2L}{n} \qquad \text{wavelength of} \\ \text{the nth harmonic} \tag{37}$$

The easiest way to remember Equation 36 is draw a sketch of the allowed standing waves and write down the progression  $\lambda_1 = 2L/1$ ,  $\lambda_2 = 2L/2$ , etc.

## **Frequency of Guitar String Waves**

What notes do you hear when you pluck a guitar string? That depends very much upon how you pluck it. Usually when you pluck the string, you create a number of the standing wave patterns at one time, and the note you hear is a rich mixture of the frequencies of the individual waves.

With care, however, you can pluck the string so that most of the vibration is in one of the harmonics. A gentle pluck at the center of the string will excite mostly the fundamental or first harmonic. Pluck the string 1/4 of the way from one end and briefly place your finger at the center of the string to create a node there. This way you can excite mostly the second harmonic. You will then notice that the sound of the note is one octave above the sound of the fundamental. Accomplished guitar players can selectively excite still higher harmonics.

What are the frequencies of oscillations of these various standing wave patterns? We can answer this question because of our knowledge that standing waves can be made from two traveling waves moving through each other. The resulting standing wave has the same frequency and wavelength as the traveling wave, thus we can use the traveling wave formulas to determine the frequency of the standing wave.

Using dimensions, we see that a traveling wave of wavelength  $\lambda$  cm/cycle, traveling at a speed v cm/sec, has a frequency f cycles/second given by

$$f \frac{\text{cycles}}{\text{sec}} = v \frac{\text{meter}}{\text{sec}} \times \frac{1}{\lambda \text{ meter/cycle}}$$
$$= \frac{v}{\lambda} \frac{\text{cycles}}{\text{sec}}$$
(38)

The speed v of a transverse wave on a string, that has a mass per unit length  $\mu$ , and tension T, is from Equation 5

$$v_{wave} = \sqrt{\frac{T}{\mu}}$$
(5)

Using Equation 5 in Equation 38 gives us as the formula for the frequency of the traveling wave

$$f = \frac{1}{\lambda} \sqrt{\frac{T}{\mu}}$$
(39)

The same Equation 39 must also apply to the standing waves on the guitar string. We get for the frequency  $f_n$  of the nth harmonic, which has a wavelength  $f_n$ ,

$$f_{n} = \frac{1}{\lambda_{n}} \sqrt{\frac{T}{\mu}} \qquad frequency of the \\ nth harmonic \qquad (40)$$

For anyone who has tuned a guitar, Equation 40 makes a lot of sense. First note that when you go from the first harmonic  $\lambda_1 = 2L$ , to the second harmonic  $\lambda_2 = L$ , the wavelength is cut in half and the frequency doubles. A doubling of the frequency of a note corresponds to going up one octave.

When you are tuning the guitar, you raise the frequency of a string by tightening it and increasing the tension. That is also predicted by Equation 40. On a guitar or most stringed instruments the low notes are played on fat wires that have a greater mass per unit length  $\mu$  than the skinny wires used for the high notes. The reason for using the fat wire is that you can increase the tension and still keep the frequency down. The more tension in the wire and the more mass in the wire, the more energy you can store in the wire and the louder the sound you can produce. It is hard to get as much sound out of the low frequency strings and you need all the help you can get.

#### **Exercise 8**

You have a guitar string of length L, with a tension T and a mass per unit length  $\mu$ . (L is the distance from the nut to the bridge.)

(a) What is the f frequency of the fundamental mode of vibration? Express your answer in terms of L, T, and  $\mu$ .

(b) Show that the nth harmonic has a frequency n times as great as the fundamental.

## Exercise 9

One end of a wire is attached to a post as shown in Figure (16). The wire is then run over a pulley where a mass m is hung on the other end. The distance d from the post to the pulley is 1 meter and the mass of one meter length of the wire is 5 grams.

(a) How big a mass m must be hung on the wire in order to get the wire to vibrate in its fundamental mode at a frequency of 440 cycles/second, which is middle A? (Answer: 395.10 kg).

(b) Describe four distinct ways one could double the frequency of oscillation of the wire.

(c) How much mass would you have hung on the wire in Figure (16) to get the wire to oscillate in its fundamental mode at a frequency two octaves above middle A? (Answer 6321.63 kg.)

## **Sound Produced by a Guitar String**

When you pluck a guitar string, the standing wave on the string produces a traveling sound wave in the air. This is analogous to plucking the end of a rope to produce a traveling wave along the rope as illustrated in Figure (8b). The wavelength of the sound wave is determined by the frequency of oscillation of the string and the speed of sound. (It is not the same as the wavelength of standing waves on the string.)

## Exercise 10

m

A guitar string is tuned to oscillate at a frequency of 440 cycles/second in its fundamental mode. What are the wavelengths of the sound waves produced by the first three harmonics

(a) in air at 20° C

(b) in helium at 20° C



# Chapter 16 Fourier Analysis, Normal Modes and Sound

In Chapter 15 we discussed the principle of superposition—the idea that waves add, producing a composite wave that is the sum of the component waves. As a result, quite complex wave structures can be built from relatively simple wave forms. In this chapter our focus will be on the analysis of complex wave forms, finding ways to determine what simple waves went into constructing a complex wave.

As an example of this process, consider what happens when white sunlight passes through a prism. White light is a mixture of all the colors, all the wavelengths of the visible spectrum. When white light passes through a prism, a rainbow of colors appears on the other side. The prism separates the individual wavelengths so that we can study the composition of the white light. If you look carefully at the spectrum of sunlight, you will observe certain dark lines; some very specific wavelengths of light are missing in light from the sun. These wavelengths were absorbed by elements in the outer atmosphere of the sun. By noticing what wavelengths are missing, one can determine what chemical elements are in the sun's atmosphere. This is how the element Helium (named after Helios, Greek for sun) was discovered.

This example demonstrates how the ability to separate a complex waveform (in this case white sunlight) into its component wavelengths or frequencies, can be a powerful research tool. The sounds we hear, like those produced by an orchestra, are also a complex mixture of waves. Even individual instruments produce complex wave forms. Our ears are very sensitive to these wave forms. We can distinguish between a note played on a Stradivarius violin and the same note played by the same person on an inexpensive violin. The only difference between the two sounds is a slight difference in the mixture of the component waves, the harmonics present in the sound. You could not tell which was the better violin by looking at the waveform on an oscilloscope, but your ear can easily tell.

You can hear these subtle differences because the ear is designed in such a way that it separates the complex incoming sound wave into its component frequencies. The information your brain receives is not what the shape of the complex sound wave is, but how much of each component wave is present. In effect, your ear is acting like a prism for sound waves.

When we study sound in the laboratory, the usual technique is to record the sound wave amplitude using a microphone, and display the resulting waveform on an oscilloscope or computer screen. If you want to, you can generate more or less pure tones that look like sine waves on the screen. Whistling is one of the best ways to do this. But if you record the sound of almost any instrument, you will not get a sine wave shape. The sound from virtually all instruments is some mixture of different frequency waves. To understand the subtle differences in the quality of sound of different instruments, and to begin to understand why these differences occur, you need to be able to decompose the complex waves you see on the oscilloscope screen into the individual component waves. You need something like an ear or a prism for these waves.

A way to analyze complex waveforms was discovered by the French mathematician and physicist Jean Baptiste Fourier, who lived from 1786 to 1830. Fourier was studying the way heat was transmitted through solids and in the process discovered a remarkable mathematical result. He discovered that any continuous, repetitive wave shape could be built up out of harmonic sine waves. His discovery included a mathematical technique for determining how much of each harmonic was present in any given repetitive wave. This decomposition of an arbitrary repetitive wave shape into its component harmonics is known as **Fourier analysis**. We can think of Fourier analysis effectively serving as a mathematical prism.

The techniques of Fourier analysis are not difficult to understand. Appendix A of this chapter is a lecture on Fourier analysis developed for high school students with no calculus background (explicitly for my daughter's high school physics class). To apply Fourier analysis you have to be able to determine the area under a curve, a process known in calculus as integration. While the idea of measuring the area under a curve is not a difficult concept to grasp, the actual process of doing this, particularly for complex wave shapes, can be difficult. Everyone who takes a calculus course knows that integration can be hard. The integrals involved in Fourier analysis, particularly the analysis of experimental data are much too hard to do by hand or by analytical means.

People find integration hard to do, but computers don't. With a computer one can integrate any experimental wave shape accurately and rapidly. As a result, Fourier analysis using a computer is very easy to do. A particularly fast way of doing Fourier analysis on the computer was discovered by Cooley and Tukey in the 1950s. Their computer technique or algorithm is known as the **Fast Fourier Transform** or **FFT** for short. This algorithm is so commonly used that one often refers to a Fourier transform as an FFT.

The ability to analyze data using a computer, to do things like Fourier analysis, has become such an important part of experimental work that older techniques of acquiring data with devices like strip chart recorders and stand alone oscilloscopes have become obsolete. With modern computer interfacing techniques, important data is best recorded in a computer for display and analysis. We have developed the MacScope<sup>TM</sup> program, which will be used often in this text, for recording and displaying experimental data. The main reason for writing the program was to make it a simple and intuitive process to apply Fourier analysis. In this chapter you will be shown how to use this program. With the computer doing all the work of the analysis, it is not necessary to know the mathematical processes behind the analysis, the steps are discussed in the appendix. But a quick reading of the appendix should give you a feeling for how the process works.

#### HARMONIC SERIES

We begin our discussion with a review of the standing waves on a guitar string, shown in Figure (15-15) reproduced here. We saw that the wavelengths  $\lambda_n$  of the allowed standing waves are given by the formula

$$\lambda_n = \frac{2L}{n} \tag{15-37}$$

Where L is the length of the string, and n takes on integer values n = 1, 2, 3, ....

Each of these standing waves has a definite frequency of oscillation that was given in Equation 15-38 as

$$f_{n}\frac{cycles}{sec} = \frac{v}{\lambda_{n}}\frac{meters/sec}{meters/cycle}$$

$$= \frac{v}{\lambda_{n}}\frac{cycles}{sec}$$
(15-38)

If we substitute the value of  $\lambda_n$  from Equation 15-37 into Equation 15-38, we get as the formula for the corresponding frequency of vibration

$$f_n = \frac{v_{wave}}{\lambda_n} = \frac{v_{wave}}{2L/n} = n \left(\frac{v_{wave}}{2L}\right)$$
(1)

For n = 1, we get

$$f_1 = \frac{v_{wave}}{2L}$$
(2)

All the other frequencies are given by

$$f_n = n f_1$$
 harmonic series (3)

This set of frequencies is called a *harmonic series*. The *fundamental frequency* or *first harmonic* is the frequency  $f_1$ . The *second harmonic*  $f_2$  has twice the frequency of the first. The third harmonic  $f_3$  has a frequency three times that of the first, etc. Note also that the fundamental has the longest wavelength, the second harmonic has half the wavelength of the fundamental, the third harmonic one third the wavelength, etc.

It was Fourier's discovery that any continuous repetitive wave could be built up by adding together waves from a harmonic series. The correct harmonic series is the one where the fundamental wavelength  $\lambda_1$  is equal to the period over which the waveform repeats.

To begin our discussion of Fourier analysis and the building up of waveforms from a harmonic series, we will first study the motion of two air carts connected by springs and riding on an air track. We will see that these *coupled air carts* have several distinct *modes of motion*. Two of the modes of motion are purely sinusoidal, with precise frequencies. But any other kind of motion appears quite complex.

However, when we record the complex motion, we discover that the velocity of either cart is repetitive. A graph of the velocity as a function of time gives us a continuous repetitive wave. According to Fourier's theorem, this waveform can be built up from sinusoidal waves of the harmonic series whose fundamental frequency is equal to the repetition frequency of the wave. When we use Fourier analysis to see what harmonics are involved in the motion, we will see that the apparent complex motion of the carts is not so complex after all.



**Figure 15-15** *Standing waves on a guitar string.* 

## NORMAL MODES OF OSCILLATION

The reason musical instruments generally produce complex sound waves containing various frequency components is that the instrument has various ways to undergo a resonant oscillation. Which resonant oscillations are excited with what amplitudes depends upon how the instrument is played.

In Chapter 14 we studied the resonant oscillation of a mass suspended from a spring, or equivalently, of a cart on an air track with springs attached to the end of the cart, as shown in Figure (1). This turns out to be a very simple system—there is only one resonant frequency, given by  $\omega = \sqrt{k/m}$ . The only natural motion of the mass is purely sinusoidal at the resonant frequency. This system does not have the complexity found in most musical instruments.



Figure 1 Cart and springs on an air track.

## Figure2

System of coupled air carts.

## Figure 3

Recording the velocity of one of the aircarts. A 10 turn coil is mounted on top of one of the carts. The coil moves through the magnetic field between the angle irons, and produces a voltage proportional to the velocity of the cart. This voltage is then recorded by the Macintosh oscilloscope. Things become more interesting if we place two carts on the air track, connected by springs as shown in Figure (2). We will call this a system of two *coupled* air carts.

When we analyzed the one cart system, we found that the force on the cart was simply F = -kx, where x is the displacement of the cart from its equilibrium position. With two carts, the force on one cart depends not only on the position of that cart, but also on how far away the other cart is. A full analysis of this coupled cart system, using Newton's second law, leads to a pair of coupled differential equations whose solution involves matrices and eigenvalues. In this text we do not want to get into that particular branch of mathematics. Instead we will study the motion of the carts experimentally, and find that the motion, which at first appears complex, can be explained in simple terms.

In order to record the motion of the aircarts, we have mounted the velocity detector apparatus shown in Figure (3). The apparatus consists of a 10 turn wire coil mounted on top of the cart, that moves through the magnetic field of the iron bars suspended above the coil. The operation of the velocity detector apparatus depends upon Faraday's law of induction which will be discussed in detail in Chapter 30 on Faraday's law. For now all we need to know is that a voltage is induced in the wire coil, a voltage whose magnitude is proportional to the velocity of the cart. This voltage signal


from the wire coil is carried by a cable to the Macintosh oscilloscope where it is displayed on the computer screen.

When you first start observing the motion of the coupled aircarts, it appears chaotic. One cart will stop and reverse direction while the other is moving toward or away from it, and there is no obvious pattern. But after a while, you may discover a simple pattern. If you pull both carts apart and let go in just the right way, the



Figure4a

Vibrational motion of the coupled air carts.



## Figure 4b

Vibrational mode of oscillation of the coupled aircarts. The voltage signal is proportional to the velocity of the cart that has the coil on top.

carts come together and go apart as if one cart were the mirror image of the other. This motion of the carts is illustrated in Figure (4a). We will call this the *vibra-tional mode* of motion.

In Figure (4b) we have used the velocity detector to record the motion of one of the coupled air carts when the carts are moving in the vibrational mode. You can see that the curve closely resembles a sine wave.





3.6



10.8

14.4

7.2

-375 -500

18.0

Seconds

If you play around with the carts for a while longer, you will discover another way to get a simple sinusoidal motion. If you pull both carts to one side, get the positions just right, and let go, the carts will move back and forth together as illustrated in Figure (5a). We will call this the *sloshing mode* of motion of the coupled air carts. In Figure (5b) we have recorded the velocity of one of the carts in the sloshing mode, and see that the curve is almost sinusoidal.

In general, the motion of the two carts is not sinusoidal. For example, if you pull one cart back and let go, you get a velocity curve like that shown in Figure (6). If you start the carts moving in slightly different ways you get differently shaped curves like the one seen in Figure (7). Only the vibrational and sloshing modes result in sinusoidal motion, all other motions are more complex. To study the complex motion of the carts, we will use the techniques of *Fourier analysis*.

## 

**Figure 6** *Complex motion of the coupled air carts.* 

## **FOURIER ANALYSIS**

As we mentioned in the introduction, Fourier analysis is essentially a mathematical prism that allows us to decompose a complex waveform into its constituent pure frequencies, much as a prism separates sunlight into beams of pure color or wavelength. We have just studied the motion of coupled air carts, which gave us an explicit example of a relatively complex waveform to analyze. While the two carts can oscillate with simple sinusoidal motion in the vibrational and sloshing modes of Figures (4) and (5), in general we get complex patterns like those in Figures (6) and (7). What we will see is that, by using Fourier analysis, the waveforms in Figures (6) and (7) are not so complex after all.

The MacScope program was designed to make it easy to perform Fourier analysis on experimental data. The MacScope tutorial gives you considerable practice using MacScope for Fourier analysis. What we will do here is discuss a few examples to see how the program, and how Fourier analysis works. We will then apply Fourier analysis to the curves of Figures (6) and (7) to see what we can learn. But first we will see how MacScope handles the analysis of more standard curves like a sine wave or square wave.



**Figure 7** Another example of the complex motion of the coupled air carts.

## **Analysis of a Sine Wave**

In Figure (8a) we attached MacScope to a sine wave generator and recorded the resulting waveform. The sine wave generator, which is usually called a *signal generator*, is an electronic device that we will use extensively in laboratory work on the electricity part of the course. Typically the device has a dial that allows you to select the frequency of the wave, a knob that allows you to adjust the wave amplitude, and some buttons by which you can select the shape of the wave. Typically you can choose between a sine wave shape,  $\checkmark$ , a square wave shape  $\neg$ .

In Figure (8a) we have selected one cycle of the wave and see that the frequency of the wave is 1515 Hz (1.515 KHz), which is about where we set the frequency dial at on the signal generator. To get Figure (8b) we pressed the *Expand* button that appears once a section of curve has been selected. This causes the selected section of the curve to fill the whole display rectangle. The selection rectangle is obtained by holding down the mouse bottom and dragging across the desired section of the curve. The starting point of the selection rectangle can be moved by holding down the shift key while moving the mouse. The data box shows the period **T** of the selected rectangle, and the corresponding frequency **f**. If you wish the data box to remain after the selection is made, hold down the option key when you release the mouse button. This immediately gives you the **ImageGrabber<sup>TM</sup>**, which allows you to select any section of the screen to save as a PICT file for use in a report or publication.

In Figure (9), we went up to the **Analyze** menu of MacScope and selected *Fourier Analysis*. As a result we get the window shown in Figure (10). At the top we see the selected one cycle of a sine wave. Beneath, we see a rectangle with one vertical bar and a scale labeled *Harmonics*. The vertical bar is in the first position, indicating that the section of the wave which we selected has only a first harmonic. Beneath the ex-



#### Figure 8a

Sine wave from a signal generator. Selecting one cycle of the wave, we see that the frequency of the wave is about 1.5 kilocycles (1.515 KHz). (Here we also see some of the controls that allow you to study the experimental data. The scrollbar labeled "S" lets you move the curve sideways. The "T" scrollbar changes the time scale, and the "O" scrollbar moves the curve up and down. In the text, we will usually not show controls unless they are important to the discussion.)



#### **Figure 8b**

Once we have selected a section of curve, a new control labeled Expand appears. When we press the Expand button, the selected section of the curve fills the entire display rectangle as seen above. (The control then becomes a Reset button which takes us back to the full





panded curve, there is a printout of the period and frequency of the selected section. Here we see again that the frequency is 1.515 KHz.

For Figure (11), we pressed **Reset** to see the full curve, selected 4 cycles of the sine wave, and expanded that. In the lower rectangle we now see one vertical bar over the 4th position, indicating that for the selected wave we have a pure fourth harmonic. In the MacScope tutorial we give you a MacScope data file for a section of a sine wave. Working with this data file, you should find that if you select one cycle of the wave, anywhere along the wave, you get an indication of a first harmonic. Select two cycles, and you get an indication of a pure second harmonic, etc.



The basic function of the Fourier analysis program in MacScope is to determine how you can construct the selected section of the wave from harmonic sine *waves*. The first harmonic has the frequency of the selected section. For example, suppose that you had a 10 cycle per second sine wave and selected one cycle. That selection would have a frequency of 10 Hz and a period of 0.1 seconds. The second harmonic would have a frequency of 20 Hz, and the third harmonic 30 Hz, etc. The nth harmonic frequency is n times greater than the first harmonic. Fourier's discovery was that any continuous repeating wave form can be constructed from the harmonic sine waves. So far we have considered only the obvious examples of sections of a pure sine wave. We will now go on to more complex examples to see how a waveform can be constructed by adding up the various harmonics.

**Figure 9** *Choosing Fourier Analysis.* 



## Figure 10

In the MacScope program, Fourier Analysis acts only on the selected section of the curve. Since we selected only one cycle, we see only a first harmonic.



## Figure 11

When 4 cycles are selected, the wave form consists of a pure 4th harmonic. We see the period of the selected section of wave, which is 4 times longer than one cycle. This makes the frequency 1/4 as high. (The difference between 1515 Hz/4 and 390.6 Hz indicates the accuracy of graphical selection.)

## Analysis of a Square Wave

The so called square wave \_\_\_\_\_, whose shape is shown in Figure (12), is commonly used in electronics labs to study the response of various electronic circuits. The waveform regularly jumps back and forth between two levels, giving it a repeated rectangular shape. The ideal mathematical square wave jumps instantaneously from one level to another. The square waves we study in the lab are not ideal; some time is always required for the transition.

It is traditional to use the square wave as the first example to show students how a complex wave form can be constructed from harmonic sine waves. This is a bit ironic, because the ideal square wave has discontinuous jumps from one level to another, and therefore does not satisfy Fourier's theorem that any *continuous wave shape* can be made from harmonic sine waves. The result is that if you try to construct an ideal square wave from sine waves, you end up with a small blip at the discontinuity (called the Gibb's effect). Since our focus is experimental data where there is no true discontinuity, we will not encounter this problem.

In Figure (12) we have selected one cycle of the square wave. Selecting *Fourier Analysis*, we get the result shown in Figure (13). We have clicked on the *Expand* button so that only the selected section of the wave



#### Figure 12

The square wave. The wave shape goes back and forth periodically between two levels. Here we have selected one cycle of the square wave.

shows in the upper rectangle. In the lower rectangle, which we will now call the FFT window, we see that this section of the square wave is made up from various harmonics. The MacScope program calculates 128 harmonics, but we have clicked three times on the *Scale* button to expand the harmonics scale so that we can study the first 16 harmonics in more detail.

In Figure (14) we have clicked on the first bar in the FFT window, the bar that represents the amplitude of the first harmonic. In the upper window you see one cycle of a sine wave superimposed upon the square wave. This is a picture of the first harmonic. It represents the best possible fit of the square wave by a single sine wave. If you want a better fit, you have to add in more sine waves.

In Figure (15) we clicked on the bar in the 3rd harmonic position, the bar representing the amplitude of the 3rd harmonic in the square wave. In the upper window you see a sine wave with a smaller amplitude and three times the frequency of the first harmonic. If you select a single harmonic, as we have just done in Figure (15), MacScope prints the frequency of both the first harmonic and the selected harmonic above the FFT window. Here you can see that the frequency of the first harmonic is 201.6 Hz and the selected harmonic frequency is 604.9 Hz as expected.



Expanding the one cycle selected, and choosing Fourier Analysis, we see that this wave form has a number of harmonics. The computer program calculates the first 128 harmonics. We used the Scale button to display only the first 16.



#### Figure 14





## Figure 15

Select the second harmonic by clicking on the second bar.



**Figure 16** Sum of the first and second harmonic is obtained by selecting both.

If we select both the first and third harmonics together (either by dragging a rectangle over both bars, or by holding down the shift key while selecting them individually), the upper window displays the sum of the first and third harmonic, as shown in Figure (16). You can see that the sum of these two harmonics gives us a waveform that is closer to the shape of the square wave than either harmonic alone. We are beginning to build up the square wave from sine waves.

In Figures (17, 18 and 19), we have added in the 5th, 7th and 9th harmonics. You can see that the more harmonics we add, the closer we get to the square wave.

One of the special features of a square wave is that it contains only odd harmonics—all the even harmonics are absent. Another is that the amplitude of the nth harmonic is 1/n times as large as that of the first harmonic. For example, the third harmonic has an amplitude only 1/3 as great as the first. This is represented in the FFT window by drawing a bar only 1/3 as high as that of the first bar. In the MacScope program, the harmonic with the greatest amplitude is represented by a bar of height 1. All other harmonics are represented by proportionally shorter bars.



## **Repeated Wave Forms**

Before we apply the Fourier transform capability of MacScope to the analysis of experimental data, there is one more feature of the analysis we need to discuss. What we are doing with the program is reconstructing a selected section of a waveform from harmonic sine waves. Anything we build from harmonic sine waves *exactly repeats at the period of the first harmonic*. Thus our reconstructed wave will always be a repeating wave, beginning again at the same height as the beginning of the previous cycle.

You can most easily see what we mean if you select a non repeating section of a wave. In Figure (20) we have gone back to a sine wave, but selected one and a half cycles. In the FFT window you see a slew of harmonics. To see why these extra harmonics are present, we have in Figure (21) selected the first 9 of them and in the upper window see what they add up to. It is immediately clear what has gone wrong. The selected harmonics are trying to reconstruct a repeating version of our 1.5 cycle of the sine wave. The extra spurious harmonics are there to force the reconstructed wave to start and stop at the same height as required by a repeating wave.

If you are analyzing a repeating wave form and select a section that repeats, then your harmonic reconstruction will be accurate, with no spurious harmonics. If



**Figure 20** When we select one and a half cycles of a sine wave, we get a whole bunch of spurious harmonics.

your data is not repeating then you have to deal one way or another with this problem. One technique often used by engineers is to select a long section of data and smoothly force the ends of the data to zero so that the selected data can be treated as repeating data. Hopefully, forcing the ends of the data to zero does not destroy the information you are interested in. You can often accomplish the same thing by throwing away higher harmonics, assuming that the lower harmonics contain the interesting features of the data. You can see that neither of these techniques would work well for our one and a half cycles of a sine wave selected in Figure (21).

In this text, our use of Fourier analysis will essentially be limited to the analysis of repeating waveforms. As long as we select a section that repeats, we do not have to worry about the spurious harmonics.

(Most programs for the acquisition and analysis of experimental data have an option for doing Fourier analysis. Unfortunately, few of them allow you to select a precise section of the experimental data for analysis. As a result the analysis is usually done on a non repeating section of the data, which distorts the resulting plot of the harmonic amplitudes. For a careful analysis of data, the ability to precisely select the data to be analyzed is an essential capability.)



The spurious harmonics result from the fact that the reconstructed wave must be repeating—must start and stop at the same height.

## ANALYSIS OF THE COUPLED AIR CART SYSTEM

We are now ready to apply Fourier analysis to our system of coupled air carts. Recall that there were two modes of motion that resulted in a sinusoidal oscillation of the carts, the vibrational motions shown in Figure (4) and the sloshing mode shown in Figure (5). In Figures (22) and (23) we expanded the time scales so that we could accurately measure the period and frequency of these oscillations. From the data rectangles, we see that the frequencies were 1.11Hz and 0.336 Hz for the vibrational and sloshing modes respectively.



Figure 22

Vibrational mode of Figure (4). We have expanded the time scale so that we could accurately measure the period and frequency of the oscillation.



Sloshing mode of Figure (5). Less of an expansion of the time scale was needed to measure the period here.

When the carts were released in an arbitrary way, we generally get the complex motion seen in Figures (6) and (7). What we wish to do now is apply Fourier analysis to these waveforms to see if any simple features underlie this complex motion.

In Figure (24), which is the waveform of Figure (6), we see that there is a repeating pattern. The fact that the pattern repeats means that it can be reconstructed from harmonic sine waves, and we can use our Fourier analysis program to find out what the component sine waves are.

In Figure (24) we have selected precisely one cycle of the repeating pattern. This is the crucial step in this experiment—finding the repeating pattern and selecting one cycle of it. How far you have to look for the pattern to repeat depends upon the mass of the carts and the strength of the springs.



Complex mode of the coupled aircarts. We see that the waveform repeats, and have selected one cycle of the repeating wave.

Expanding the repeating section of the complex waveform, and choosing *Fourier Analysis* gives us the results shown in Figure (25). What we observe from the Fourier analysis is that the complex waveform is a mixture of two harmonics, in this case the third and tenth harmonic. If we click on the bar showing the amplitude of the third harmonic, we see that harmonic



#### Figure 25

Fourier analysis of one cycle of the complex waveform. The FFT rectangle shows us that the wave consists of only two harmonics.



Figure 26

When we click on the third harmonic bar, we see that the frequency of the third harmonic is precisely the frequency of the sloshing mode of oscillation.

drawn in the display window of Figure (26), and we find that the frequency of this harmonic is 0.336 Hz. This is the frequency of the sloshing mode of Figure (23). Clicking on the bar above the tenth harmonic, we get the harmonic drawn in the display window of Figure (27), and see that the frequency of this mode is 1.12 Hz, within a fraction of a percent of the 1.11 Hz frequency of the vibrational mode of Figure (22).



#### Figure 27

Selecting the tenth harmonic, we see that its frequency is essentially equal to the frequency of the vibrational mode of motion.



Selecting both modes shows us that the complex motion is simply the sum of the two sinusoidal modes of motion.

If we select both the third and tenth harmonics, the sum of these two harmonics is shown in Figure (28). These two harmonics together so closely match the experimental data that we had to move the experimental curve down in order to see both curves. What we have learned from this experiment is that *the complex motion of Figure (6) is a mixture of the two simple, sinusoidal modes of motion.* 

Back in Figures (6) and (7), we displayed two waveforms, representing different complex motions of the same two carts. Starting with the second waveform of Figure (7), we selected one cycle of the motion, expanded the selected section, and chose Fourier Analysis. The result is shown in Figure (29). What we see is that the second complex waveform is also a mixture of the third and tenth harmonics. The first waveform in Figure (25) had more of the third harmonic, more of the sloshing mode, while the second waveform of Figure (29) has more of the tenth harmonic, the vibrational mode. Both complex waveforms are simply mixtures of the vibrational and sloshing modes. *They have different shapes because they are different mixtures*.

This experiment is beginning to demonstrate that for the two coupled aircarts, there is a strict limitation to the kind of motion the carts can have. They can either move in the vibrational mode, or in the sloshing mode,



Figure 29

The second complex mode of motion, from Figure (7), is simply a different mixture of the same vibrational and sloshing modes of motion.

## or in some combination of the two modes. No other kinds of motion are allowed! The various complex motions are just different combinations of the two modes.

Adding another aircart so that we have three coupled aircarts, the motion becomes still more complex. However if we look carefully, we find that the waveform eventually repeats. Selecting one repeating cycle and choosing Fourier Analysis, we got the results shown in Figure (30). We observe that this complex motion is made up of three harmonics.

The sinusoidal modes of motion of the coupled air carts are called *normal modes*. The general rule is that if you have n coupled objects, like n carts on an air track connected by springs, and they are confined to move in 1 dimension, there will be n normal modes. (With 2 carts, we saw 2 normal modes. With 3 carts, 3 normal modes, etc.) This result, which will play an important role in our discussion of the specific heat of molecules, can be extended to motion in 2 and 3 dimensions. For example, if you have n coupled particles that can move in 3 dimensions, as in the case of a molecule with n atoms, then the system should have 3n normal modes of motion. Such a molecule should have 3n independent ways to vibrate or move. We will have more to say about this subject in the next chapter.



Figure 30

One cycle of the waveform for three coupled aircarts. With three carts, we get three normal modes of motion.

## THE HUMAN EAR

The human ear performs a frequency analysis of sound waves that is not unlike the Fourier analysis of wave motion which we just studied. In the ear, the initial analysis is done mechanically, and then improved and sharpened by a sophisticated data analysis network of nerves. We will focus our attention on the mechanical aspects of the ear's frequency analysis.

Figure (31) is a sketch of the outer and inner parts of the human ear. Sound waves, which consist of pressure variations in the air, are funneled into the auditory canal by the external ear and impinge on the eardrum, a large membrane at the end of the auditory canal. The eardrum (tympanic) membrane vibrates in response to the pressure variations in the air. This vibrational motion is then transferred via a lever system of three bones (the malleus, incus, and stapes) to a small membrane covering the oval window of the snail shaped cochlea.

The cochlea, shown unwound in Figure (32), is a fluid filled cavity surrounded by bone, that contains two main channels separated by a membrane called the *basilar membrane*. The upper channel (scala vestibuli) which starts at the oval window, is connected at the far end to the lower channel (scala tympani) through a hole called the helicotrema. The lower channel returns to the round window which is also covered by a membrane. If the stapes pushes in on the membrane at the oval window, fluid flows around the helicotrema and causes a bulge at the round window.



The purpose of the lever system between the eardrum and the cochlea is to efficiently transfer sound energy to the cochlea. The eardrum membrane is about 25 times larger in area than the membrane across the oval window. The lever system transfers the total force on the eardrum to an almost equal force on the oval window membrane. Since force equals pressure times area, a small pressure variation acting on the large area of the eardrum membrane results in a large pressure variation at the small area at the oval window. The higher pressures are needed to drive a sound wave through the fluid filled cochlea.

If the oval window membrane is struck by a pulse, a pressure wave travels down the cochlea. The basilar membrane, which separates the two main fluid channels, moves in response to the pressure wave, and a series of hair cells along the basilar membrane detect the motion. It is the way in which the basilar membrane responds to the pressure wave that allows for the frequency analysis of the wave.

Figure (33) is an idealized sketch of a straightened out cochlea. (See Appendix B for more realistic sketches.) At the front end, by the oval window, the basilar

membrane is narrow and stiff, while at the far end it is about 5 times as wide and much more floppy. To see why the basilar membrane has this structure, we have in Figure (34) sketched a mechanical model that has a similar function as the membrane. In this model we have a series of masses mounted on a flexible steel band and attached by springs to fixed rods as shown. The masses are small and the springs stiff at the front end. If we shake these small masses, they resonate at a high frequency  $\omega = \sqrt{k/m}$ . Down the membrane model, the masses get larger and the springs weaken with the result that the resonant frequency becomes lower.

If you gently shake the steel band at some frequency  $\omega_0$ a small amplitude wave will travel down the band and soon build up a standing wave of that frequency. If  $\omega_0$ is near the resonant frequency of one of the masses, that mass will oscillate with a greater amplitude than the others. Because the masses are connected by the steel band, the neighboring masses will be carried into a slightly larger amplitude of motion, and we end up with a peak in the amplitude of oscillation centered around the mass whose frequency  $\omega$  is equal to  $\omega_0$ . (In the sketch, we are shaking the band at the resonant frequency  $\omega_7$  of the seventh mass.)



#### Figure 34

Spring model of the basilar membrane. As we go down the steel band, the masses become larger, the springs weaker, and the resonant frequency drops. If we vibrate the end of the band at some frequency, the mass which resonates at that frequency will have the biggest amplitude of oscillation.

Because the masses m get larger and the spring constants k get smaller toward the far end, there is a continuous decrease in the resonant frequency as we go down the band. If we start shaking at a high frequency, the resonant peak will occur up near the front end. As we lower the frequency, the peak of the oscillation will move down the band, until it finally gets down to the lowest resonant frequency mass at the far end. We can thus measure the frequency of the wave by observing where along the band the maximum amplitude of oscillation occurs.

The basilar membrane functions similarly. The stiff narrow membrane at the front end resonates at a high frequency around 20,000 Hz, while the wide floppy back end has a resonant frequency in the range of 20 to 30 Hz. Figure (35) shows the amplitude of the oscillation of the membrane in response to driving the fluid in the cochlea at different frequencies. We can see that as the frequency increases, the location of the maximum amplitude moves toward the front of the membrane, near the stapes and oval window. Figure (36) depicts the shape of the membrane at an instant of maximum amplitude when driven at a frequency of a few hundred Hz. The amplitude is greatly exaggerated; the basilar membrane is about 33 millimeters long and the amplitude of oscillation is less than .003 mm.

Although the amplitude of oscillation is small, it is accurately detected by a system of about 30,000 hair cells. How the hair cells transform the oscillation of the membrane into nerve impulse signals is discussed in Appendix B at the end of this chapter.

The human ear is capable of detecting tiny changes in frequency and very subtle mixtures of harmonics in a sound. Looking at the curves in Figure (35) (which were determined from a cadaver and may not be quite as sharp as the response curves from a live membrane), it is clear that it would not be possible to make the ear's fine frequency measurements simply by looking for the peak in the amplitude of the oscillation of the membrane. But the ear does not do that. Instead, measurements are continuously made all along the membrane, and these results are fed into a sophisticated data analysis network before the results are sent to the brain. The active area of current research is to figure out how this data analysis network operates.



#### Figure 35

Amplitude of the motion of the basilar membrane at different frequencies. (Adapted from Principles of Neural Science Edited by E. R. Kandel and J. H. Schwartz, Elsevier/North-Holland, p 263.)



#### Figure 36

Response of the basilar membrane to a moderately low frequency driving force. (From Vander,A; Sherman,J; and Luciano,D. Human Physiology, 4th edition, 1985, P662. McGraw Hill Publishing Co., NY.)

## STRINGED INSTRUMENTS

The stringed instruments provide the clearest example of how musical instruments function. The only possible modes of oscillation of the string are those with nodes at the ends, and we have seen that the frequencies of these modes form a harmonic series  $f_n = n f_1$ . This suggests that if we record the sound produced by a stringed instrument and take a Fourier transform to see what harmonics are present in the sound, we can tell from that what modes of oscillation were present in the vibrating string.

This is essentially correct for the electric stringed instruments like the electric guitar and electric violin. Both of these instruments have a magnetic pickup that detects the velocity of the string at the pickup, using the same principle as the velocity detector we used in the air cart experiments discussed earlier in this chapter. The voltage signal from the magnetic pickup is then amplified electronically and sent to a loudspeaker. Thus the sound we hear is a fairly accurate representation of the motion of the string, and an analysis of that sound should give us a good idea of which modes of oscillation of the strings were excited.

The situation is different for the acoustic stringed instruments, like the acoustic guitar used by folk singers, and the violin, viola, cello and base, found in symphony orchestras. In these instruments the vibration of the string does not produce that much sound itself. Instead, the vibrating string excites resonances in the sound box of the instrument, and it is the sound produced by the resonating sound box that we hear. As a result the quality of the sound from an acoustic string instrument depends upon how the sound box was constructed. Subtle differences in the shape of the sound box and the stiffness of the wood used in its construction can lead to subtle differences in the harmonics excited by the vibrating string. The human ear is so sensitive to these subtle differences that it can easily tell the difference between a great instrument like a 280 year old Stadivarius violin, and even the best of the good instruments being made today. (It may be that it takes a couple of hundred years of aging for a very good violin to become a great one.)

To demonstrate the difference between electric and acoustic stringed instruments, and to illustrate how Fourier analysis can be used to study these differences, my daughter played the same note, using the same bowing technique, on the open E string of both her electric and her acoustic violins. Using the same microphone in the same setting to record both, we obtained the results shown in Figure (37).

From Figure (37) we immediately see why acoustic stringed instruments sound differently from their electric counterparts. With the acoustic instruments you get a far richer mixture of harmonics. In the first trial, labeled E(1) Electric Violin, the string was bowed so that it produced a nearly pure 4th harmonic. The sound you hear corresponds to a pure tone of frequency 657.8/4 = 264 Hz. The corresponding sound produced by the acoustic violin has predominately the same 4th harmonic, but a lot of the sound is spread through the first 8 harmonics.

A number of recordings were made, so that we could see how the sounds varied from one playing to the next. The examples shown in Figure (37) are typical. It is clear in all cases that for this careful bowing a single harmonic predominated in the electric violin while the acoustic violin produced a mixture of the first eight. It is rather surprising that the ear hears all of these sounds as representing the same note, but with a different quality of sound.

We chose the violin for this comparison because by using a bow, one can come much closer to exciting a pure mode of vibration of the string. We see this explicitly in the E(1) Electric Violin example. When you pluck or strum a guitar, even if you pluck only one string, you get a far more complex sound than you do for the violin. If you pluck a chord on an acoustic guitar, you get a very complex sound. It is the complexity of the sound that gives the acoustic guitar a richness that makes it so effective for accompanying the human voice.

## Figure 37

Comparison of the sound of an electric and an acoustic violin. In each case the open E string was bowed as similarly as possible. The electric violin produces relatively pure tones. The interaction between the string and the sound box of the acoustic violin gives a much richer mixture of harmonics.



Recording the sound of an instrument and using Fourier analysis can be an effective tool for studying musical instruments, but care is required. For example, in comparing two instruments, start by choosing a single note, and try to play the note the same way on both instruments. Make several recordings so that you can tell whether any differences seen are due to the way the note was played or due to the differences in the instruments themselves. With careful work, you can learn a lot about the nature of the instrument.

In the 1970s, before we had personal computers, students doing project work would analyze the sound produced by instruments or the voice, using a time sharing mainframe computer system to do the Fourier analysis. Hours of work were required to analyze a single sound, but the results were so interesting that they served as the incentive to develop the MacScope<sup>TM</sup> program when the Macintosh computer became available

I particularly remember an early project in which two students compared the spinet piano, the upright piano and the grand piano. They recorded middle C played on each of these pianos. Middle C on the spinet consisted of a wide band of harmonics. From the upright there were still a lot of harmonics, but the first, third, and fifth began to predominate. The grand piano was very clean with essentially only the first, third, and fifth present. You could clearly see the effect of the increase in the size of the musical instrument.

The same year, another student, Kelly White, took a whale sound from the Judy Collins record *Sound of the Humpback Whales*. Listening to the record, the whale sounds are kind of squeaky. But when the sound was analyzed, the results were strikingly similar to those of the grand piano. The analysis suggested that the whale sounds were by an instrument as large as, or larger, than a grand piano. (The whale's blowhole acts as an organ pipe when the whale makes the sound.)

## WIND INSTRUMENTS

While the string instruments are all based on the oscillation of a string, the wind instruments, like the organ, flute, trumpet, clarinet, saxophone, and glass bottle, are all based on the oscillations of an air column. Of these, the bottle is the most available for studying the nature of the oscillations of an air column.

When you blow carefully across the top of a bottle, you hear a sound with a very definite frequency. Add a little water to the bottle and the frequency of the note rises. When you shortened the length of a string, the pitch went up, thus it is not surprising that the pitch also goes up when you shorten the length of the air column.

You might guess that the mode of vibration you set up by blowing across the top of the bottle has a node at the bottom of the bottle and an anti node at the top where you are blowing. For a sine wave the distance from a node to the next anti node is 1/4 of a wavelength, thus you might predict that the sound has a wavelength 4 times the height of the air column, and a frequency

$$f = \frac{v_{sound}}{\lambda} = \frac{v_{sound}}{4d}$$
 (for  $\lambda = 4d$ ) (4)

This prediction is not quite right as you can quickly find out by experimenting with various shaped bottles. Add water to different shaped bottles, adjusting the levels of the water so that all the bottles have the same height air columns. When you blow across the top of the different bottles, you will hear distinctly different notes, the fatter bottles generally having lower frequencies than the skinny ones. Unlike the vibrating string, it is not just the length of the column and the speed of the wave that determine the frequency of oscillation, the shape of the container also has a noticeable effect.

## Figure 38

Blowing across the top of a bottle produces a note whose wavelength is approximately 4 times the height of the air column.



Despite this additional complexity, there is one common feature to the air columns encountered in musical instruments. They all have unique frequencies of oscillation. Even an organ pipe that is open at both ends—a situation where you might think that the length of the column is not well defined—the air column has a precise set of frequencies. The fussiness in defining the length of the column does not fuzz out the sharpness of the resonance of the column.

Organ pipes, with their straight sides, come closest to the simple standing waves we have seen for a stretched instrument string. In all cases, the wave is excited at one end by air passing over a sharp edge creating a turbulent flow behind the edge. This turbulence excites the air column in much the same way that dragging a sticky bow across a violin string excites the oscillation of the string.



**Figure 39** b) Modes of oscillation of an air column in open and closed organ pipes.

The various modes of oscillation of an air column in an organ pipe are shown in Figure (39). All have an anti node at the end with sharp edges where the turbulence excites the oscillation. The open ended pipes shown in Figure (39a) also have anti nodes at the open end, while the closed pipes of Figure (39b) have a node at the far end. The pictures in Figure (39a) are a bit idealized, but give a reasonably accurate picture of the shape of the standing wave. We can compensate for a lack of accuracy of these pictures by saying, for example, that the anti node of the open ended pipes lies somewhat beyond the end of the pipe.

## Exercise 1

(a) Find the formula for the wavelength  $\lambda_n$  of the nth harmonic of the open ended pipes of Figure (39).

(b) Assuming that the frequencies of vibration are given by  $f(cycle/sec) = v(meter/sec) / \lambda(meter/cycle)$  where v is the speed of sound, what is the formula for the allowed frequencies of the open organ pipe of length L?

(c) What should be the length L of an open ended organ pipe to produce a fundamental frequency of 440 cycles/ second, middle A?

(d) Repeat the calculations of parts a, b, and c for the closed organ pipes of Figure (39).

(e) If you have the opportunity, find a real organ and check the predictions you have made (or try the experiment with bottles).

If you start with an organ pipe, and drill holes in the side, essentially converting it into a flute or one of the other wind instruments like a clarinet, you considerably alter the shape and frequency of the modes of oscillation of the air column. As a first approximation you could say that you create an anti node at the first open hole. But then when you play these instruments you can make more subtle changes in the pitch by opening some holes and closing others. The actual patterns of oscillation can become quite complex when there are open holes, but the simple fact remains that, no matter how complex the wave pattern, there is a precise set of resonant frequencies of oscillation. It is up to the maker of the instrument to locate the holes in such a way that the resonances have the desired frequencies.

## PERCUSSION INSTRUMENTS

We all know that the string and wind instruments produce sound whose frequency we adjust to produce melodies and chords. But what about drums? They seem to just make noise. Surprisingly, drumheads have specific modes of oscillation with definite frequencies, just as do vibrating strings and air columns. But one does not usually adjust the fundamental frequency of oscillation of the drumhead, and the frequencies of the higher modes of oscillation do not follow the harmonic patterns of string and wind instruments.

To observe the standing wave patterns corresponding to modes of vibration of a drumhead, we can drive the drumhead at the resonant frequency of the oscillation we wish to study. It turns out to be a lot easier to drive a drumhead at a precise frequency than it is to find the normal modes of the coupled air cart system.

The experiment is illustrated in Figure (40). The apparatus consists of a hollow cardboard cylinder with a rubber sheet stretched across one end to act as a drumhead. At the other end is a loudspeaker attached to a signal generator. When the frequency of the signal generator is adjusted to the resonant frequency of one of the normal modes of the drumhead, the drumhead will start to vibrate in that mode of oscillation.

To observe the shape and motion of the drumhead in one of its vibrational modes, we place a strobe light to one side of the drumhead as shown. If you adjust the strobe to the same frequency as the normal mode vibration, you can stop the motion and see the pattern.



**Figure 40** Studying the modes of oscillation of a drumhead.



Figure 41

Modes of oscillation of a drumhead. (Adapted from Vibration and Sound by Phillip M. Morse, 2nd ed., McGraw-Hill, New York, 1948.)

Turn the frequency a bit off resonance, and you get a slow motion moving picture of the motion of that mode.

Some of the low frequency normal mode or standing wave patterns of the drumhead are illustrated in Figure (41). In the lowest frequency mode, Figure (41a), the entire center part of the drumhead moves up and down, much like a guitar string in its lowest frequency mode. In this pattern there are no nodes except at the rim of the drumhead.

In the next lowest frequency mode, shown in Figure (41b), one half the drumhead goes up while the other half goes down, again much like the second harmonic mode of the guitar string. The full two dimensional nature of the drumhead standing waves begins to appear in the next mode of Figure (41c) where the center goes up while the outside goes down. Now we have a circular node about half way out on the radius of the drumhead.

As we go up in frequency, we observe more complex patterns for the higher modes. In Figure (41d) we see a pattern that has a straight node like (41b) and a circular node like (41c). This divides the drumhead into 4 separate regions which oscillate opposite to each other. Finer division of the drumhead into smaller regions can be seen in Figures (41e) and (41f). The frequencies of the various modes are listed with each diagram. You can see that there is no obvious progression of frequencies like the harmonic progression for the modes of a stretched string. When you strike a drumhead you excite a number of modes at once and get a complex mixture of frequencies. However, you do have some control over the modes you excite. Bongo drum players, for example, get different sounds depending upon where the drum is struck. Hitting the drum in the center tends to excite the lowest mode of vibration and produces a lower frequency sound. Striking the drum near the edges excites the higher harmonics, giving the drum a higher frequency sound.

Even more complex than the modes of vibration of a drumhead are those of the components of a violin. To construct a successful violin, the front and back plates of a violin must be tuned before assembly. Figure (42) shows a violin backplate under construction, while Figure (43) shows the first 6 modes of oscillation of a completed backplate. Note again that the resonant frequencies do not form a harmonic series.

Back plate of a violin under construction. The resonant frequencies are tuned by carving away wood from different sections of the plate.





Figure 43

Modes of oscillation of the backplate made visible by holographic techniques. Quality violins are made by tuning the frequencies of the various modes. (Figures 42 and 43 from "The Acoustics of Violin Plates," by Carleen Maley Hutchins, Scientific American, October 1981.)

## SOUND INTENSITY

One of the amazing features of the human eardrum is its ability to handle an extreme range of intensities of sound waves. We define the *intensity* of a sound wave as the *amount of energy per second being carried by a sound wave through a unit area*. In the MKS system of units, this would be the number of joules per second passing through an area of one square meter. Since one joule per second is a unit of power called a *watt*, the MKS unit for sound intensity is *watts per square meter*. The human ear is capable of detecting sound intensities as faint as  $10^{-12}$  watts/m<sup>2</sup>, but can also handle intensities as great as 1 watt/m<sup>2</sup> for a short time. This is an astounding range, a factor of  $10^{12}$  in relative intensity.

The ear and brain handle this large range of intensities by essentially using a logarithmic scale. Imagine, for example, you are to sit in front of a hi fi set playing a pure tone, and you are told to mark off the volume control in equal steps of loudness. The first mark is where you just barely hear the sound, and the final mark is where the sound just begins to get painful. Suppose you are asked to divide this range of loudness into what you perceive as 12 equal steps. If you then measured the intensity of the sound you would find that the intensity of the sound increased by approximately a factor of 10 after each step. Using the faintest sound you can hear as a standard, you would measure that the sound was 10 times as intense at the end of the first step, 100 times as intense after the second step, 1000 times at the third, and  $10^{12}$  times as intense at the final step.

The idea that the intensity increases by a factor of 10 for each equal step in loudness is what we mean by the statement that the loudness is based on a logarithmic scale. We take the faintest sound we can hear, an intensity  $I_0 = 10^{-12}$  watts/m<sup>2</sup> as a basis. At the first setting  $I = I_0$ , at the second setting  $I_1 = 10 I_0$ , at the 3rd setting  $I_2 = 100 I_0$ , etc. The factor I/I<sub>0</sub> by which the intensity has increased is thus

$$I_0/I_0 = 1 \qquad (I_0 = 10^{-12} \text{ watts/m}^2)$$

$$I_1/I_0 = 10$$

$$I_2/I_0 = 100$$

$$\dots$$

$$I_{12}/I_0 = 10^{12} \qquad (5)$$

10

Taking the logarithm to the base 10 of these ratios gives

$$Log_{10}(I_0/I_0) = Log_{10}(1) = 0$$
  

$$Log_{10}(I_1/I_0) = Log_{10}(10) = 1$$
  

$$Log_{10}(I_2/I_0) = Log_{10}(100) = 2$$
  
.....  

$$Log_{10}(I_{12}/I_0) = Log_{10}(10^{12}) = 12$$
(6)

## **Bells and Decibels**

The scale of loudness defined as  $\text{Log}_{10}(\text{I}/\text{I}_0)$  with  $\text{I}_0 = 10^{-12}$  watts/m<sup>2</sup> is measured in bells, named after Alexander Graham Bell, the inventor of the telephone.

$$\left| \begin{array}{c} \text{loudness of a sound} \\ \text{measured in bells} \end{array} \right\rangle \equiv \text{Log}_{10} \left( \frac{\text{I}}{\text{I}_0} \right) \right|$$

$$I_0 = 10^{-12} \text{watts/m}^2$$
(7)

From this equation, we see that the faintest sound we can hear, at  $I = I_0$ , has a loudness of zero bells. The most intense one we can stand for a short while has a loudness of 12 bells. All other audible sounds fall in the range from 0 to 12 bells.

It turns out that the bell is too large a unit to be convenient for engineering applications. Instead one usually uses a unit called the *decibel (db)* which is 1/10 of a bell. Since there are 10 decibels in a bell, the formula for the *loudness*  $\beta$ , in decibels, is

$$\beta$$
(decibels) = (10 db)Log<sub>10</sub> $\left(\frac{I}{I_0}\right)$  (8)

On this scale, the loudness of sounds range from 0 decibels for the faintest sound we can hear, up to 120 decibels  $(10 \times 12 \text{ bells})$  for the loudest sounds we can tolerate.

The average loudness of some of the common or well known sounds is given in Table 1.

Table 1Various Sound Levels in db	
threshold of hearing	0
rustling leaves	10
whisper at 1 meter	20
city street, no traffic	30
quiet office	40
office, classroom	50
normal conversation at 1 meter	60
busy traffic	70
average factory	80
jack hammer at 1 meter	90
old subway train	100
rock band	120
jet engine at 50 meters	130
Saturn rocket at 50 meters	200

An increase in loudness of 10 db corresponds to an increase of 1 bell, or an increase of intensity by a factor of 10. A rock band at 110 db is some 100 times as intense (2 factors of 10) as a jack hammer at 90 db. A Saturn rocket is about  $10^{20}$  times as intense as the faintest sound we can hear.

Our sensitivity to sound depends not only to the intensity of the sound, but also to the frequency. About the lowest frequency note one can hear, and still perceive as being sound, is about 20 cycles/second. As you get older, the highest frequencies you can hear decreases from around 20,000 cycles/sec for children, to 15,000 Hz for young adults to under 10,000 Hz for older people. If you listen to too much, too loud rock music, you can also decrease your ability to hear high frequency sounds.

Figure (44) is a graph of the average range of sound levels for the human ear. The faintest sounds we can detect are in the vicinity of 4000 Hz, while any sound over 120 db is almost uniformly painful. The frequency ranges and sound levels usually encountered in music are also shown.



## Figure 44

Average range of sound levels for the human ear. Only the very young can hear sound frequencies up to 20,000 Hz. (Adapted from Fundamental Physics by Halliday and Resnick, John Wiley & Sons.)

## **Sound Meters**

Laboratory experiments involving the intensity or loudness of sound are far more difficult to carry out than those involving frequencies like the Fourier analysis experiments already discussed. From the output of any reasonably good microphone, you can obtain a relatively good picture of the frequencies involved in a sound wave. But how would you go about determining the intensity of a sound from the microphone output? (There are commercial sound meters which have a scale that shows the ambient sound intensity in decibels. Such devices are often owned by zoning boards for checking that some factory or other noise source does not exceed the level set by the local zoning ordinance, often around 45 db. The point of our question is, how would you calibrate such a device if you were to build one?)

The energy in a wave is generally proportional to the square of the amplitude of the wave. A sound wave can be viewed as oscillating pressure variations in the air, and the energy in a sound wave turns out to be proportional to the square of the amplitude of the pressure variations. The output of a microphone is more or less proportional to the amplitude of the pressure variations, thus we expect that the intensity of a sound wave should be more or less proportional to the square of the voltage output of the microphone. However, there is a great variation in the sensitivity of different microphones, and in the amplifier circuits used to produce reasonable signals. Thus any microphone that you wish to use for measuring sound intensities has to be calibrated in some way.

Perhaps the easiest way to begin to calibrate a microphone for measuring sound intensities is to use the fact that very little sound energy is lost as sound travels out through space. Suppose you had a speaker radiating 100 watts of sound energy, and for simplicity let us assume that the speaker radiates uniformly in all directions and that there are no nearby walls.

If we are 1 meter from this speaker, all the sound energy is passing out through a 1 meter radius sphere centered on the speaker. Since the area of a sphere is  $4\pi r^2$ , this 1 meter radius sphere has an area of  $4\pi$  meters<sup>2</sup>, and the average intensity of sound at this 1 meter distance must be

average intensity of  
sound 1 meter from  
a 100 watt speaker 
$$= \frac{100 \text{ watts}}{4\pi \text{ meters}^2}$$
$$= 8.0 \frac{\text{watts}}{\text{meters}^2}$$
(9)

If we wish to convert this number to decibels, we get

$$\beta \begin{pmatrix} \text{sound intensity 1 meter} \\ \text{from a 100 watt speaker} \end{pmatrix}$$
  
= (10 db) Log  $\frac{I}{I_0}$   
= (10 db) Log\_{10}  $\left( \frac{8.0 \text{ watts / m}^2}{10^{-12} \text{ watts / m}^2} \right)$   
= (10 db) × Log\_{10} (8 × 10^{12})  
= (10 db) × 12.9  
= 129 db (10)

From our earlier discussion we see that this exceeds the threshold of pain. One meter from a 100 watt speaker is too close for our ears. But we could place a microphone there and measure the amplitude of the signal output for our first calibration point. Move the speaker back to a distance of 10 meters and the area that the sound energy has to pass through increases by a factor of 100 since the area of a sphere is proportional to  $r^2$ . Thus as the same 100 watts passes through this 100 times larger area, the intensity drops to 1/100 of its value at 1 meter. At a distance of 10 meters the intensity is thus 8/100 = .08 watts/m<sup>2</sup> and the loudness level is

$$\beta \begin{pmatrix} 10 \text{ meters from a} \\ 100 \text{ watt speaker} \end{pmatrix}$$
$$= (10 \text{ db}) \text{Log}_{10} \left( \frac{.08 \text{ watts / } \text{m}^2}{10^{-12} \text{ watts / } \text{m}^2} \right)$$
$$= 109 \text{ db}$$
(11)

We see that when the intensity drops by a factor of 100, it drops by 20 db or 2 bells.

To calibrate your sound meter, record the amplitude of the signal on your microphone at this 10 meter distance, then set the microphone back to a distance of 1 meter, and cut the power to the speaker until the microphone reads the same value as it did when you recorded 100 watts at 10 meters. Now you know that the speaker is emitting only 1/100th as much power, or 1 watt. Repeating this process, you should be able to calibrate a fair range of intensities for the microphone signal. If you get down to the point where you can just hear the sound, you could take that as your value of I<sub>0</sub>, which should presumably be close to I<sub>0</sub> =  $10^{-12}$ watts/m<sup>2</sup>. Then calibrate everything in db and you have built a loudness meter. (The zoning board, however, might not accept your meter as a standard for legal purposes.)

## Exercise 2

What is the loudness, in db, 5 meters from a 20 watt speaker? (Assume that the sound is radiated uniformly in all directions).

## **Exercise 3**

You are playing a monophonic record on your stereo system when one of your speakers cuts out. How many db did the loudness drop? (Assume that the intensity dropped in half when the speaker died. Surprisingly you can answer this question without knowing how loud the stereo was in the first place. The answer is that the loudness dropped by 3 db).

## Speaker Curves

When you buy a hi fi loudspeaker, you may be given a frequency response curve like that in Figure (45), for your new speaker. What the curve measures is the intensity of sound, at a standard distance, for a standard amount of power input at different frequencies. It is a fairly common industry standard to say that the frequency response is "flat" over the frequency range where the intensity does not fall more than 3 db from its average high value. In Figure (45), the response of that speaker, with the woofer turned on, is more or less "flat" from 62 Hz up to 30,000 Hz.

Why the 3 db cutoff was chosen, can be seen in the result of Exercise 3. There you saw that *if you reduce the intensity of the sound by half, the loudness drops by 3 db*. This is only 3/120 (or 1/36) of our total hearing range, not too disturbing a variation in what is supposed to be a flat response of the speaker.





Speaker response curve from a recent audio magazine. The dashed line shows the response when the woofer is turned off. (We added the dotted lines at + and -3 db.)

# **APPENDIX A** FOURIER ANALYSIS LECTURE

In our discussion of Fourier Analysis, we saw that any wave form can be constructed by adding together a series of sine and cosine waves. You can think of the Fourier transform as a mathematical prism which breaks up a sound wave into its various wavelengths or frequencies, just as a light prism breaks up a beam of white light into its various colors or wavelengths. In MacScope, the computer does the calculations for us, figuring out how much of each component sine wave is contained in the sound wave. The point of this lecture is to give you a feeling for how these calculations are done. The basic ideas are easy, only the detailed calculations that the computer does would be hard for us to do.

## **Square Wave**

In Figure A-1 we show a MacScope window for a square wave produced by a Hewlett Packard oscillator.

We have selected precisely one cycle of the wave, and see that the even harmonics are missing. A careful investigation shows that the amplitude of the Nth odd harmonic is 1/N as big as the first (e.g., the 3rd harmonic is 1/3 as big as the first, etc.). Thus the mathematical formula for a square wave F(t) can be written:

$$F(t) = (1)\sin(t) + (1/3)\sin(3t) + (1/5)\sin(5t) + (1/7)\sin(7t) ...$$

where, for now, we are assuming that the period of the wave is precisely  $2\pi$  seconds. The coefficients (1), (1/3), (1/5), (1/7), which tell us how much of each sine wave is present, are called the *Fourier coefficients*. Our goal is to calculate these coefficients.

## **Calculating Fourier Coefficients**

In general we cannot construct an arbitrary wave out of just sine waves, because sine waves, sin(t), sin(2t), etc., all have a value 0 at t = 0 and at  $t = 2\pi$ . If our wave is not zero at the beginning (t = 0) of our selected period, or not zero at the end  $(t=2\pi)$ , then we must also include cosine waves which have a value 1 at those points. Thus the general formula for breaking an arbitrary repetitive wave into sine and cosine waves is:

$$F(t) = A_0 + A_1 \cos(1t) + A_2 \cos(2t) + A_3 \cos(3t) + ... + B_1 \sin(1t) + B_2 \sin(2t) + B_2 \sin(3t) + ... (A-1)$$

The question is: How do we find the coefficients  $A_0$ ,  $A_1$ ,  $A_2$ ,  $B_1$ ,  $B_2$  etc. in Equation (A1)? (These are the *Fourier coefficients.*)

To see how we can determine the Fourier coefficients, let us take an explicit example. Suppose we wish to find the coefficient  $B_3$ , representing the amount of sin(3t) present in the wave. We can find  $B_3$  by first multiplying Equation (1) through by sin(3t) to get:



**Figure A-1** *The square wave has only odd harmonics.* 

$$F(t) \sin(3t) = A_0 \sin(3t) + A_1 \cos(1t) \sin(3t) + A_2 \cos(2t) \sin(3t) + A_3 \cos(3t) \sin(3t) + ... + B_1 \sin(1t) \sin(3t) + B_2 \sin(2t) \sin(3t) + B_3 \sin(3t) \sin(3t) + ... (A-2)$$

At first it looks like we have created a real mess. We have a lot of products like cos(t)sin(3t), sin(t)sin(3t), sin(3t)sin(3t), etc. To see what these products look like, we plotted them using **True BASIC**<sup>TM</sup> and obtained the results shown in Figure (A2) (on the next page).

Notice that in all of the plots involving sin(3t), the product  $sin(3t)sin(3t) = sin^2(3t)$  is special; it is the only one that is always positive. (It has to be since it is a square.) A careful investigation shows that, in all the other non square terms, there is as much negative area as positive area, as indicated by the two different shadings in Figure (A3). If we define the "net area" under a curve as *the positive area minus the negative area*, then only the sin(3t)sin(3t) term on the right side of the Equation A-2 has a net area.

The mathematical symbol for finding the net area under a curve (in the interval t = 0 to  $t = 2\pi$ ) is:



**Figure A-3** Only the square terms such as sin(3t)sin(3t)have a net area. This curve has no net area.

Area under  

$$\sin(3t)\sin(3t)$$
  
from t = 0  
to t =  $2\pi$ 

$$= \int_{0}^{2\pi} \sin(3t)\sin(3t)dt$$
 (A-3)

(Those who have had calculus say we are taking the **integral** of the term sin(3t)sin(3t).)

A basic rule learned in algebra is that if we do the same thing to both sides of an equation, the sides will still be equal. This is also true if we do something as peculiar as evaluating the net area under the curves on both sides of an equation.

If we take the net area under the curves on the right side of Equation A-2, only the  $\sin^2(3t)$  term survives and we get:

$$\int_{0}^{2\pi} F(t)\sin(3t)dt = B_3 \int_{0}^{2\pi} \sin^2(3t)dt \qquad (A-4)$$

In Figure (A4), we have replotted the curve  $\sin^2(3t)$ , and drawn a line at height y = .5. We see that the peaks above the y = .5 line could be flipped over to fill in the valleys below the y = .5 line. Thus  $\sin^2(3t)$  in the interval t = 0 to  $t = 2\pi$  has a net area equal to that of a rectangle of height .5 and length  $2\pi$ . I.e., the net area is  $\pi$ :

$$\int_{0}^{2\pi} \sin^2(3t) dt = \pi$$
 (A-5)



**Figure A-4** The area under the curve  $\sin^2(3t)$  is equal to the area of a rectangle .5 high by  $2\pi$  long.



2π

2π

2π

2π

2π

**Figure A-2** *Product wave patterns.* 

Substituting Equation A-5 in Equation A-4 and solving for the Fourier coefficient  $B_3$  gives:

$$B_3 = \frac{1}{\pi} \int_0^{2\pi} F(t) \sin(3t) dt$$
 (A-6)

Similar arguments show that the general formulas for the Fourier coefficients  $A_n$  and  $B_n$  are:

$$A_n = \frac{1}{\pi} \int_0^{2\pi} F(t) \cos(nt) dt$$
 (A-7)

$$B_{n} = \frac{1}{\pi} \int_{0}^{2\pi} F(t) \sin(nt) dt$$
 (A-8)

These integrals, which were nearly impossible to do before computers, are now easily performed even on small personal computers. Thus the computer has made Fourier analysis a practical experimental tool.

## **Amplitude and Phase**

Instead of writing the Fourier series as a sum of separate sine and cosine waves, it is often more convenient to use amplitudes and phases. The basic formula we use is \*\*

$$A\cos(t) + B\sin(t) = C\cos(t-\phi)$$
 (A-9)

where



\*\* Start with

 $\begin{array}{l} \cos(x-y) = \ \cos(x) \times \cos(y) + \sin(x) \times \sin(y) \\ \text{Let } x = t, \ y = \phi, \ \text{and multiply through by } C \ \text{to get} \\ C \cos(t-\phi) = \ \{C \cos(\phi)\} \cos(t) + \{C \sin(\phi)\} \sin(t) \\ \text{This is Equation } A-9 \ \text{if we set} \\ A = \{C \cos(\phi)\}; \ B = \{C \sin(\phi)\} \\ \text{Thus } \ \tan\phi = \ \sin\phi/\cos\phi = \ B/A \end{array}$ 

The function  $\cos(t-\phi)$  is illustrated in Figure (A-5). We see that C is the amplitude of the wave, and the phase angle  $\phi$  is the amount the wave has been moved to the right. (When t=0,  $\cos(t-\phi) = \cos(-\phi)$ .) With Equation A-9, we can rewrite equation A-1 in the form:

$$F(t) = C_0 + C_1 \cos(t - \phi_1) + C_2 \cos(2t - \phi_2) + C_3 \cos(3t - \phi_3) + ...$$
(A-10)

The advantage of Equation A-10 is that the coefficients C represent how much of each wave is present, and sometimes we do not care about the phase angle  $\phi$ . For example our ears are not particularly sensitive to the phase of the harmonics in a musical note, thus the tonal quality of a musical instrument is determined almost entirely by the amplitudes C of the harmonics the instrument produces.



Figure A-5

The function  $\cos(t-\phi)$ . When  $\phi = 90^{\circ}$  you get a sine wave.

In the Fourier transform plots we have shown so far, the graph of the harmonics has been representing the amplitudes C. If you wish to see a plot of the phases  $\phi$ , then press the button labeled ø as shown, and you get the result seen in Figure(A-6). In that figure we are looking at the phases of the odd harmonic sine waves that make up a square wave. Since

$$\sin(t) = \cos(t - 90^\circ)$$

all the sine waves should have a phase shift of  $90^{\circ}$ .

If for any reason, you need accurate values of the Fourier coefficients, they become available if you press the **FFT Data** button to get the results shown in Figure (A-7). When you do this, the **Editor** window is filled with a text file containing the A, B coefficients accurate to 3 or 4 significant figures.



## Figure A-6

When you press the  $\phi$  button, the Fourier Transform display shifts from amplitudes to phases. Since the square wave is made up of pure odd harmonic sine waves, each odd harmonic should have a 90 degree or  $\pi/2$  phase shift.

		Fourier Trar	nsform 🔳		
$F(T) = \Sigma[A]$	nCos(n*T) +	BnSin(n*T)	)]		ĥ
Amplitude	= J[An(T)^2	+ Bn(T)^2	]		888
Phase = In	vTan(Bn/An)	(degrees)			
Amplitudes	scaled so	that large:	st is 1.00		
*					
Harmonic	Amplitude	Phase	An(T)	Bn(T)	
1	1.00	91	0115	. 7027	
2	.00	0	.0043	.0000	
3	.33	91	0026	. 2329	
4	.00	0	.0043	.0000	
5	.19	88	.0047	.1403	
6	.00	0	.0043	.0000	$\overline{\mathbb{C}}$
7	.14	89	.0013	.1009	먼
P			ſ		
.5-			ł	Scale ]	
	_		ŀ		
البال (ف	╷┨╷凮╷		╸╷╺╸╷┤		
1	4 8	12	16	Harmonic	

## Figure A-7

For greater numerical accuracy, you can press the FFT Data button. This gives you a text file with the A, B coefficients given to four places.



#### **Figure A-8**

Amplitude. The Fourier transform of a 3-slit pattern gives the amplitude of the diffraction pattern that would be produced by a laser beam passing through these slits. (Selecting the data to give wider slits would correspond to using a different wavelength laser beam.)



#### Figure A-9

Intensity. In the lab, you see the intensity of the diffraction pattern. MacScope will display the intensity of the Fourier transform if you click one more time on the ø button.

## Amplitude and Intensity

An experiment that has become possible with Mac-Scope, is to have students compare the Fourier transform of a multiple slit grating with the diffraction pattern produced by a laser beam passing through that grating. For example, in Figure (A-8) we have taken the transform of a 3-slit grating. In this case, the 3-slit "pattern" was made simply by turning a 2 volt power supply on and off. We are now working on ways for students to record the slit pattern directly.

The problem with Figure (A-8) is that the Fourier transform of the slits gives the amplitude of the diffraction pattern, while in the lab one measures the intensity of the diffraction pattern. The intensity of a light wave is proportional to the square of its amplitude.

In order that students can compare a slit Fourier transform with an experimental diffraction pattern, we have designed MacScope so that one more press on the ø button takes us from a display of phases to intensities. Explicitly, the ø button cycles from *amplitudes* to *phases* to *intensities*. In Figure (A-9) we have clicked over to intensities, and this pattern may be directly compared with the intensity of the 3-slit diffraction pattern seen in lab.

(The Fourier transform of the diffraction pattern amplitude should give the slit pattern. Unfortunately, if you take the Fourier transform of the experimental diffraction pattern, you are taking the transform of the intensity, or square, of the amplitude. What you get, as Chris Levey of our department demonstrated, is the *convolution* of the slit pattern with itself.)

# **APPENDIX B** INSIDE THE COCHLEA

In Figure (32), our simplified unwound view of the cochlea, we show only the basilar membrane separated by two fluid channels (the scala vestibuli which starts at the oval window, and the scala tympani which ends at the round window). That there is much more structure in the cochlea is seen in the cochlea cross section of Figure (B-1). The purpose of this additional structure is to detect the motion of the basilar membrane in a way that is sensitive to the harmonic content of the incoming sound wave.

Recall that when the basilar membrane is excited by a sinusoidal oscillation, the maximum amplitude of the response of the basilar membrane is located at a position that depends upon the frequency of the oscillation.



Figure 32 (repeated) The cochlea unwound.



## Figure B-1

Cross section of the cochlea. (From Vander,A; Sherman,J; and Luciano,D. Human Physiology, 4th edition, 1985, P662. McGraw Hill Publishing Co., NY.)

As seen in Figure (35), the lower the frequency, the farther down the membrane the maximum amplitude occurs. Along the top of the basilar membrane is a system of hair cells that detects the motion of the membrane and sends the needed information to the brain.

Figure (B-2a) is a close up view of the hair cells that sit atop of the basilar membrane. (There are about 30,000 hair cells in the human ear.) Above the hair cells is another membrane called the tectorial membrane which is hinged on the left hand side of that figure. Fine hairs go from the top of each hair cell up to the tectorial membrane as shown. When the basilar membrane is deflected by an incoming sound wave the hairs are bent as shown in Figure (B-2b). It is the bending of the hairs that triggers an electrical impulse in the hair cell.

Figure (B-3) is a mechanical model of how the bending of the hairs creates the electrical impulse. The fluid in the cochlea duct surrounding the hair cells has a high concentration of positive potassium ions  $(k^+)$ . The



## Figure 35 (repeated)

Amplitude of the motion of the basilar membrane at different frequencies, as we go down the basilar membrane. bending of the hair cell opens small channels allowing potassium ions to flow into the hair cell. This flow of positive charge into the cell changes the electrical potential of the cell, triggering reactions that will eventually result in an electrical impulse in the nerve fiber that is connected to the hair cell.

After the channel at the top of the hair cell closes, the excess potassium is pumped out of the hair cell, and the cell returns to its normal resting voltage, ready to fire again.

There are various ways that a hair cell can transmit frequency information to the nervous system. One is by its location down the basilar membrane. The lower the frequency of the sound wave, the farther down the membrane an oscillation of the membrane takes place. Thus high frequency waves excite cells at the front of the basilar membrane, while low frequency oscillations excite cells at the back end.

Secondly, hair calls in a given area show special sensitivities to different frequencies. Figure (B-4) shows

## Figure B-3

Model of the valves at the base of the hair cells. (From Shepard,G.M., Neurobiology, 3rd Edition, 1994, P316. Oxford Univ. Press.)



the amplitude, in db, of the sound wave required to excite a nerve fiber connected to that particular region of hair cells. You can see that the nerve is most sensitive to a 2 killocycle (2kHz) frequency. At 2 kHz, that nerve fires when excited by a 15 db sound wave. It is not excited by a 4 kHz wave until the sound intensity rises to 80 db.

Ultimately the exquisite sensitivity of the human ear to different frequency components in a sound wave results from the fact that there are about 30,000 hair cells continuously monitoring the motion of the basilar membrane. Effective processing of this vast amount of information leads to the needed sensitivity. Much of this processing of information occurs in the nervous system in the ear, before the information is sent to the brain.

Figure B-4 Frequency dependence. A much lower amplitude sound will excite this nerve fiber at 2kHz than any other frequency. Different nerve fibers connected to the hair cells have different frequency dependence.





#### Figure B-2 a,b

When the basilar membrane is deflected by a sound wave, the hair cells are bent. This opens a valve at the base of the hair call. (Figures 2 & 4 adapted from Kandel, E; Schwartz, J; and Jessell, T; Principles of Neural Science, 3rd Edition, 1991; pages 486 and 489.)

# Chapter 17 Atoms, Molecules and Atomic Processes

To extract the basic laws of mechanics from the variety and confusion of the world around us required looking at matters on a large scale, looking out at the moon and planets whose motion is regular, periodic, and easier to understand. In this and the next two chapters we take a similarly large leap to the small scale of distance where simplicity and periodic behavior again allow us to gain insight into the working of nature. Here we find the world of atoms and their constituent particles, a world in which we observe the basic forces and particles ultimately responsible for the variety about us.

The jump down to the small scale of atoms is comparable to the jump out from the study of projectile motion in the lab to the analysis of satellite orbits. Imagine, for example, that we could enlarge the golf balls used in our strobe labs to the size of the earth. The same enlargement of a hydrogen atom would give us an object about the size of a golf ball.

Only with the development of the new generation of microscopes in the late 1980s has it become possible to see and work with individual atoms. Figure (1) is the first atomic sized logo consisting of xenon atoms on a background of nickel, made by scientists at the IBM Research Laboratories in 1990. But despite great improvements in seeing and working with individual atoms, the images we now get are still fuzzy and we are restricted to looking at atoms in solid structures, atoms that do not move around. Our knowledge of atoms comes not from looking through microscopes, but instead from the study of chemical reactions, the measurement of the physical properties of substances, and the bombardment of materials with x-rays and other particles. This study essentially began with John Dalton's construction of the first periodic table in 1808. Other milestones were Thomson's discovery of the electron in 1895, Rutherford's discovery of the atomic nucleus in 1912, Neils Bohr's model of the hydrogen atom in 1913, and the discovery of the rules of quantum mechanics in the mid 1920s.



#### Figure 1

Thirty five xenon atoms were dragged across a nickel surface to form the letters IBM. (D. M. Eigler & E. K. Schweizer, Nature, 5 April 1990.)

## MOLECULES

Atoms attract each other to form molecules, like the water molecule  $H_20$  sketched in Figure (2). It was from x-ray studies of ice, the crystalline form of water, that we know the distance from the center of the oxygen atom to the center of ahydrogen atom is  $.958 \times 10^{-8}$  cm and that the hydrogen atoms are spread out at an angle of 104.5 degrees as shown.

X-ray studies of large biological molecules began in the late 1950s. For example, myoglobin is a substance found in muscle tissue. The myoglobin molecule contains over 2500 atoms, mostly carbon, hydrogen, oxygen, nitrogen, and one iron atom. For determining the precise structure of the myoglobin molecule from x-rays of crystals of myoglobin, John Kendrew and Max Peritz received the 1963 Nobel Prize in chemistry. Their model of the molecule is shown in Figure (3).

Recent advances in computer modeling now provide detailed views of numerous kinds of molecules. An example is Figure (4) showing the cholera toxin Bsubunit.







**Figure 4** *Computer model of the cholera toxin B-subunit. (Courtesy of Argonne National Laboratory.)* 



**Figure 3** Model of the myoglobin molecule. (Photograph courtesy of J.C. Kendrew and H.C. Watson.)

## **ATOMIC PROCESSES**

The myoglobin molecule provides a hint of the complex structures that can be formed from atoms, a complexity we wish to avoid in this chapter by concentrating on basic, simpler atomic processes. To help illustrate these processes, we have illustrated in Figures (5) through (11) a set of sketches drawn on the blackboard by Richard Feynman and copied to his book of introductory physics lectures. Such simple sketches, full of information, were characteristic of Feynman's style.

In the first three Figures (5, 6, and 7) we have views of three forms of matter made from water molecules, namely ice, water and steam. In the form of ice, the water molecules fit into a hexagonal structure with a hole in the center, as seen in Figures (5a,b). When water freezes to form snowflakes, the hexagonal structure repeats and we get a six sided symmetry seen in all snowflakes, examples of which is shown in Figure (5c).

When ice melts to form water, shown in Figure (6), the rigid structure of ice disappears and the water molecules can now slide past each other. In addition the holes in the hexagonal structure fill in with the result that water is more dense than ice. That is why ice floats in water.

The third form of water is steam, shown in Figure (7). In the gaseous state the water molecules move freely about, interacting only when they collide with each other. The picture of steam is more or less what we would see if we could look at the steam emerging from a teakettle on an atomic scale. The separation of the molecules is on the average about 10 times the diameter of a water molecule. As a result, the steam is about 1000 times less dense than liquid water. The transition from water to steam, either by evaporation or by boiling, involves a competition between molecular forces and thermal forces. We will discuss that competition shortly.



**Figure 5a** Ball and stick model of an ice crystal.



Sketch of the arrangement of the water molecules in an ice crystal.



**Figure 5c** Snowflakes reflect the 6-sided structure of the ice crystal.

Some atomic processes are illustrated in Figures (8) through (10). Figure (8) is what you might see in a snapshot of the surface of a glass of water. On our scale of distance, such a surface looks quiet, but on an atomic scale it is an active place with molecules continually entering and leaving the water. Evaporation occurs if more water molecules leave the water than return. If you put a cover over the glass, the concentration of water molecules in the air above the water builds up to the point where just as many water molecules return as leave, and the level of the water stops dropping. We would then say that the evaporation has ceased.

In Figure (9) we see what happens to a block of carbon when it is heated in an atmosphere of oxygen. By themselves oxygen atoms combine in pairs to form  $O_2$  or oxygen molecules, and carbon atoms attract each



**Figure 6** *Water magnified a billion times.* 



Figure 7 Steam.

other to form solids like diamond, graphite, soot, or Buckeyballs (soccerball shaped structures of carbon). But there is a greater attraction between a carbon and an oxygen atom than between two carbon or two oxygen atoms. If the carbon and oxygen are heated, the various atoms bounce into each other at high speeds, the old structures break apart and molecules of carbon monoxide  $\bigcirc$  and carbon dioxide  $\bigcirc \bigcirc$  are formed. Energy is released in the process in the form of heat and light, and we say that the carbon is burning.



oxygen hydrogen nitrogen

**Figure 8** *Water evaporating in air.* 



O oxygen

carbon

**Figure 9** *Carbon burning in oxygen.*
The process of salt dissolving in water is illustrated in Figure (10). Table salt is a stable crystal structure made from sodium and chlorine atoms. Strong electric forces hold these atoms together in the crystal. But let water molecules come into contact with the salt crystal, and the water molecules work their way in between the sodium and chlorine atoms, allowing these atoms to move freely and independently throughout the water.

One of our favorite sketches is Figure (11), the odor of violets. Many of our common experiences have a simple origin on an atomic scale.





Figure 10 Chlorine Salt dissolving in water.



# Odor of violets.

# THERMAL MOTION

What we have not been able to display in the sketches of atomic processes is the constant juggling of the atoms and molecules. This juggling, which we will call *thermal motion*, becomes more intense as a substance becomes warmer and can cause major changes in the structure of matter. When ice is warmed to above the melting point, the juggling or thermal motion breaks up the rigid structure of the ice crystal, allowing the water molecules to slide past each other to form liquid water. With more heating, the thermal motion can increase to the point that the water molecules fly apart.

Surprisingly this thermal motion can be seen on a considerably larger scale than the scale of atoms. In 1827, the botanist Robert Brown observed that tiny pollen particles in water, when seen through a microscope, moved around in a juggling, random fashion. Wondering whether these particles were alive and swimming, Brown studied these and other small particles in circumstances where nothing should be alive, and concluded that this random motion, now called *Brownian motion*, had nothing to do with life, but was related to the motion of the molecules.

With a laser, microscope and TV camera, it is easy to set up a demonstration of the Brownian motion of cigarette smoke particles in air. The apparatus, shown in Figure (12a), consists of a small cavity between two microscopic slides, into which we inject smoke through a small tube.





Brownian Motion. A small cavity, made from microscope slides is filled with cigarette smoke, is illuminated from the side by a laser beam and viewed from above by a microscope and TV camera.

Once inside the cavity the smoke is illuminated from the side by a laser beam. Through the microscope, whose image can be displayed using a TV camera, we clearly see the illuminated smoke particles moving around in a relatively slow jagged motion. Figure (12b) is one minute movie showing the motion of the smoke particles.

In Figure (12c), a student recorded the motion of a single smoke particle for 38 TV frames. Although we may think of smoke particles as being small, they are huge compared to the air molecules in which they are immersed. Smoke particles have a mass many orders of magnitude larger than that of the air molecules. Yet the motion we see is caused by the constant bombardment of these huge particles by the air molecules.

At first you might believe that if a large particle were constantly bombarded on all sides by billions of tiny particles the effect of the collisions would cancel out and the big particle would just sit there. But it turns out that if the collisions are random, then fluctuations in the collisions will cause the large particle to move around with the jerky motion we see in the Brownian motion demonstration. What is more remarkable, *the average kinetic energy of the smoke particles, as they wander about, is the same as the average kinetic energy of the air molecules bombarding them.* 



#### Figure 12b

First frame of a one minute movie showing the brownian motion of smoke particles. The frames are 1/15 of a second apart in this movie. (Click on the image to see the movie. Press the "esc" button to close the movie.) The air molecules themselves are not all the same; air is mostly nitrogen and oxygen, some carbon dioxide and water, and smaller amounts of other gases and pollutants. It turns out that each species of molecule in the air has precisely the same average kinetic energy due to thermal motion. As a result, the oxygen molecules, for example, having a slightly greater mass than the nitrogen molecules, must have a slightly smaller average speed in order that the average kinetic energy  $1/2 \text{ mv}^2$  be the same. The smoke particles, with their huge masses, have a much slower average speed than the air molecules. The air molecules move at roughly the speed of sound in air, while the smoke particles move slowly enough for us to see and follow them on the TV screen.



## Figure 12c

Brownian motion of a smoke particle. This is the result of a student project by Lisa Stigler, where the motion of a smoke particle was recorded by a TV camera using the apparatus of Figure (12a). (We cannot use this plot to estimate the average speed of the smoke particles, because the smoke particle undergoes many collisions between TV frames. However this plot does illustrate the random walk nature of the motion of the particle. One feature of a random walk, that may be observable from plots like this, is that the average distance from the starting point should be proportional to the square root of the elapsed time.)

## **Exercise 1**

If we know the relative masses of the molecules in air, and know the average speed of one of the species, we can use the fact that all species of particles have the same average kinetic energy, in order to calculate the average speed of the other particles. It turns out that if we have a sample of air at room temperature (27° C), the average speed of the nitrogen molecule is 483 meters/ sec. (We will calculate this number shortly.)

Using the mass of a hydrogen atom as a standard mass of 1, the mass of a nitrogen molecule is 28, and an oxygen molecule 32. (These are often called the **molecular weights** of the molecules). In Table 1 we have listed various constituents of the cigarette smoke, the relative mass of the particles, and the average thermal speed of two of the species. Use the fact that all the species have the same average kinetic energy to fill in the table of average speeds. (We gave you the speed of helium so that you could check your calculations.)

Particle	Symbol	Mass*	Speed**			
Hydrogen molecule	$H_2$	2.0				
Helium atom	He	4.0	1370 m/sec			
Water molecule	$H_20$	18.0				
Nitrogen molecule	$N_2$	28.0	518 m/sec			
Oxygen molecule	$0_2$	32.0				
Carbon Dioxide	$CO_2$	44.0				
Smoke particle		$10^{10}$				
* Mass relative to Hydrogen atom						
** Average speed at room temperature						

## Table 1

Particles involved in the Brownian motion demonstration. Fill in the column for average speed, using the fact that these particles have the same average kinetic energy.

# THERMAL EQUILIBRIUM

If you place a hot cup of coffee and a cold glass of milk on the table and leave the room for several hours, when you return the coffee and the milk are both at room temperature. We say that the coffee, the milk, and the air in the room are in thermal equilibrium. A faster way to reach thermal equilibrium, at least between the coffee and the milk, is to pour the milk into the coffee and stir.

On an atomic scale, what does it mean to say that the molecules of the coffee and those of the milk are in thermal equilibrium? We obtain a hint from our discussion of Brownian motion. The cigarette smoke represents a well stirred mixture of smoke particles and air molecules. These should therefore be in thermal equilibrium, just as the well stirred molecules in the coffee and milk. In the case of Brownian motion, the smoke particles and the air molecules had the same average thermal kinetic energy. We expect that the same may be true for the molecules in the mixture of coffee and milk.

It is an almost general rule that when two objects are in thermal equilibrium, the molecules that make up these objects have the same average thermal kinetic energy. When we first placed a hot cup of coffee and a cold cup of milk on the table, the molecules in the coffee had a greater average kinetic energy, and the molecules in the milk a lesser average kinetic energy, than the molecules in the air. But after a few hours, the coffee molecules slowed down and the milk molecules speeded up until all three sets of molecules, coffee, milk, and air attained the same average kinetic energy.

The process of reaching thermal equilibrium is usually a result of random collisions between molecules. If a fast molecule collides with a slow one, chances are that the slow one will speed up and the fast one will slow down. It requires a detailed analysis, which we will not attempt, to show that if the collisions are random, they tend to equalize the kinetic energy of the particles.

# TEMPERATURE

If there is any scientific concept familiar to everybody, it is the concept of temperature. All of our lives we have been poked with thermometers and listened to weather forecasts about tomorrow's temperature. How is that quantity, measured by various kinds of thermometers, related to the atomic processes we have been discussing?

Most thermometers are a black art which depends upon such properties as the thermal expansion of mercury or alcohol, the stiffness of a spring, or the color changes of a material, etc. There is, however, one kind of a thermometer whose function can be understood from a simple molecular picture. That is the *ideal gas thermometer* which we will discuss shortly. We will see that for an ideal gas thermometer, the temperature reading is proportional to the average thermal kinetic energy of the gas molecules in the thermometer.

When you measure the temperature of an object, you have to wait until the thermometer and the object are in thermal equilibrium. (You wait until the reading on the thermometer in your mouth stops changing.) When in thermal equilibrium, the molecules of the object and those of the thermometer have the same average thermal kinetic energy. If we are using an ideal gas thermometer, the reading is proportional to this average kinetic energy. Thus if we use an ideal gas thermometer as an experimental definition of temperature, we are effectively defining temperature as being proportional to the average thermal kinetic energy of the molecules.

## Absolute Zero

An immediate consequence of temperature being related to thermal kinetic energy, is that there must be a lowest possible temperature. When the thermal kinetic energy is gone, you cannot go any lower in temperature. It thus seems reasonable to define an absolute zero of temperature as the state where the molecules have no thermal kinetic energy, and choose a temperature scale that starts at this absolute zero and goes up proportionally to the thermal kinetic energy.

However, as you approach absolute zero, as you try to remove the last vestiges of thermal kinetic energy, nature has a surprise in store. No matter what you do, there is some unremovable kinetic energy left. One of the basic predictions of quantum mechanics is that a confined particle cannot have zero kinetic energy, and the closer the confinement the more kinetic energy it has to have. A molecule in a liquid or a solid is confined to the small volume bounded by its neighbors, and therefore cannot have a kinetic energy less than that required for that volume.

The unremovable kinetic energy is called *zero point energy*. This energy is so small that for most substances it is not noticeable unless you carry out specially designed experiments to detect it. However, zero point energy shows up clearly in the case of liquid helium. All substances except helium freeze when cooled to a sufficiently low temperature. We can remove enough kinetic energy from the molecules so that they settle into a solid structure. But the molecular force between helium atoms is so weak that the zero point kinetic energy alone is enough to keep helium a liquid. You cannot freeze helium by cooling alone, you must also subject it to high pressures.

The existence of zero point energy suggests that we will encounter problems with the definition of temperature as we approach absolute zero. Suppose, for example, we have two substances with different zero point energies in thermal equilibrium. If the temperature is so low that any thermal kinetic energy is much less than the zero point energies, then we have a situation in which molecules with different vibrational kinetic energies are in thermal equilibrium. If we insist that two substances in thermal equilibrium are at the same temperature, then we can no longer say that temperature is proportional to the vibrational kinetic energy of the molecules.

The ideal gas thermometer does not get us out of this problem because it does not work at very low temperatures. Before the zero point energies become important, any gas we use in an ideal gas thermometer becomes liquid or solid and we no longer have an ideal gas as a working substance.

In the next chapter on entropy and the second law of thermodynamics, we will discuss the consequences of the basic idea that *order does not naturally arise from disorder*. In that discussion we will describe a method of defining temperature that applies to all temperature ranges. This thermodynamic definition of temperature is consistent with the ideal gas thermometer over the range that ideal gas thermometers operate, but also correctly describes temperatures near absolute zero where we have to deal with zero point energy.

# **Temperature Scales**

For the rest of this chapter, we will put aside any worries about zero point energy, and simply assume that the temperature of an object is proportional to the average thermal kinetic energy of the molecules in the object, and that absolute zero is where no thermal kinetic energy remains.

From this point of view, the simplest way to define a temperature scale is to equate the temperature with the average thermal kinetic energy, and measure temperature in energy units such as ergs as shown in Figure (13). But you probably have not heard anyone describe temperature in ergs, and for good reason. Telling your doctor that you are running a fever of  $6.4423 \times 10^{-14}$ , an increase of  $23 \times 10^{-18}$  over normal could be a bit hard to explain when you are sick. It is much easier to say that you have a temperature of  $100^{\circ}$  F or about 38° C. Ergs are too awkward a unit for most purposes.

Historically, thermometers were invented and temperature scales established long before the relation between temperature and the average kinetic energy of molecules became known. Throughout the world the most widely used temperature scale is the Centigrade scale, where the temperature of melting ice is arbitrarily set at 0° C (zero degrees Centigrade), and the



**Figure 13a** *Temperature scale in ergs.* 

**Figure 13b** *Comparison of various temperature scales.* 

boiling of water at  $100^{\circ}$  C. Commonly, changes in temperature are measured with a mercury thermometer. This device registers temperature changes when the mercury in a thin glass column expands or contacts. On the Centigrade scale, the distance between  $0^{\circ}$  C and  $100^{\circ}$  C is marked into 100 equally spaced smaller intervals which we call degrees.

A less arbitrary scale is the *Kelvin* or *absolute* scale, which measures temperature in Centigrade size degrees beginning at absolute zero. Using the absolute scale, we find that helium boils at 4 degrees Kelvin, ice melts at 273 degrees Kelvin and water boils at 373 degrees Kelvin. A comparison of various temperature scales (ergs, degrees Kelvin, degrees Centigrade, and degrees Fahrenheit) is shown in Figure (13b).

Those who define standard nomenclature for physical quantities have decided, in their great wisdom, that the word "degrees" shall be omitted when talking about temperature in degrees Kelvin. Thus we should say that helium boils at 4 kelvins or 4K, ice melts at 273 kelvins or 273K, and the temperature difference between melting ice and boiling water is 100 kelvins or 100K. At least this nomenclature is easy to say and should not be confusing when you get used to it. We do not feel the same way about all recent changes in nomenclature.

The conversion from one temperature scale to another is a relatively straightforward process. If you went to an American school, somewhere along the way you were taught how to convert from Fahrenheit to Centigrade degrees. You do not need to worry about that because we will not be using the obsolete Fahrenheit scale. But we will often want to convert from the absolute scale to the energy units, ergs or joules. The conversion is written in the somewhat peculiar form

average kinetic energy  
of gas molecules  
in ergs or joules 
$$= \frac{3}{2} kT$$
 (1)

where T is the temperature in kelvins, and the conversion factor k, known as *Boltzman's constant* has the numerical value

$$\frac{\text{Boltzman's}}{\text{constant k}} = 1.38 \times 10^{-16} \frac{\text{ergs}}{\text{kelvin}}$$
(2)

If you are using MKS units, the value of k is  $1.38 \times 10^{-23}$  joules/kelvin.

The important feature of Equation 1 is that the average kinetic energy of the molecules is proportional to the absolute temperature measured in kelvins. We have written the proportionality constant as 3/2 k, putting in the numerical factor of 3/2 to get rid of a factor 2/3, as you will see shortly. Basically, think of Boltzman's constant as the conversion factor to go from temperature units to energy units or vice versa.

## Exercise 2

Use Equation 1 to calculate the temperature of melting ice in ergs. Compare your answer with the result in Figure (13).

## Exercise 3

What would be the temperature in Kelvins of a gas if the particles in the gas had an average kinetic energy of 1 erg?

## Exercise 4

In Exercise 1 we said that the average speed of nitrogen molecules at room temperature was 518 meters/sec, and asked you to use that result to calculate the average speed of the other molecules and particles in the cigarette smoke. Now you are to calculate the speed of the nitrogen molecules using the fact that their average kinetic energy is 3/2 kT.

It is traditional to take room temperature as  $300K = 27^{\circ}$  C. Assume that a nitrogen molecule is 28 times as massive as a hydrogen atom, whose mass is essentially the same as a proton, or  $1.67 \times 10^{-24}$  grams. See if you get the answer of 518 meters/sec.

# **MOLECULAR FORCES**

Much of the behavior of matter we see as we look around us is the result of a competition between molecular forces holding atoms together and thermal motion tending to pull them apart. Molecular forces can be subtle enough to form objects as complex as the myoglobin molecule. Yet knowing just some of the basic features of molecular forces is enough to provide an insight into processes like evaporation, osmotic



## Figure 14

Interaction of two atoms via a Leonard Jones potential (f). When the atoms have an equilibrium separation (d), their potential energy of interaction is a minimum, and we can visualize one of the atoms as sitting at the bottom of the potential energy well. If the separation either increases or decreases, there is a force back toward equilibrium. The repulsion quickly builds up if you try to shove the atoms together, and the attraction dies rapidly after the atoms become separated by about one atomic diameter. pressure, elasticity of rubber, and the behavior of an ideal gas. At the end of Chapter 19 we will discuss the so-called *Leonard Jones potential* as a model for molecular forces. That model is far more detailed than we need for our current discussion. All we need to know now is reviewed in Figure (14).

In Figure (14a), where the atoms are about an atomic diameter apart, the attractive molecular force between the two atoms is less than one percent of its maximum value. The point is that unless the atoms are very close together, within an atomic diameter of each other, molecular forces are negligible. This is why atoms in a gas often act as independent free particles. In the air we breath, the average spacing of atoms is about ten molecular diameters, so that molecular forces play no role except when molecules collide.

When atoms get closer than an atomic diameter, the attractive molecular force increases rapidly, reaching a maximum at a separation at about one tenth of an atomic diameter as shown in Figure (14c). Then the force rapidly drops to zero at the spacing shown in Figure (14d). When the force is zero, this is the equilibrium distance which determines the size of the atom in a chunk of matter. Effectively we can say that when the atoms are at their equilibrium separation, they are just touching, as we drew them in Figure (14d).

Try to shove the atoms closer together than the equilibrium position, and you encounter a repulsive force that builds very rapidly, much faster than the attractive force increases as you pull the atoms apart. This repulsion makes atoms behave as hard, nearly incompressible spherical objects. This repulsive force is often referred to as the repulsive core of the atom.

In Figure (14f) we have sketched the potential energy corresponding to the molecular force. As you can see the potential energy forms a well with the bottom at the equilibrium position. When two atoms form a molecule, like hydrogen ( $H_2$ ), oxygen ( $O_2$ ) or nitrogen ( $N_2$ ), you can picture one of the atoms as sitting in the potential well created by the other, and vice versa. We only have to think about one of the atoms, for the same thing is happening to the other.

If the atom is sitting at rest at the bottom of its potential well, it is at its equilibrium position shown in Figure (14d). If the atom moves either way, in or out, it is subject to a restoring force pushing it back to the equilibrium position. If it does not move too far from its equilibrium position, the restoring force is very similar to the restoring force of a spring at equilibrium, as indicated in Figure (15). That is why one can often quite accurately picture molecular forces as spring forces between atoms.



a) Comparison of the Leonard Jones potential and the parabolic potential of a spring force.



b) Modeling the molecular force as a spring force. As long as the atom stays near the equilibrium position, the Leonard Jones force and the spring force are equivalent.

#### Figure 15

Physics and chemistry texts often picture molecules with spring forces between the atoms. At first this may seem to be a highly unrealistic picture. But when you carefully compare the spring potential energy and the Leonard Jones potential energy right near the equilibrium position, the two curves have the same shape. Thus the spring force is a good model as long as the atoms stay near their equilibrium positions. The advantage of the potential energy diagram is that it allows us to think of atomic processes in terms of the energy involved. The distance from the zero of potential energy down to the bottom of the well is the *binding energy* of the molecule as indicated in Figure (16). This is the energy required to pull a molecule apart, starting with the atoms in their equilibrium position.

We have seen that the average thermal kinetic energy of an atom or molecule is 3/2 kT where k is Boltzman's constant and T the temperature in Kelvins. If the thermal kinetic energy is much less than the binding energy of the molecule, as we have shown in Figure (16), then the atom can move back and forth around the bottom of the potential well but not climb out.

From the depth and shape of the potential well one can deduce general features of the behavior of matter, such as why solids and liquids expand when heated. But before we look at such fine details, there is much to understand about atomic processes just from the fact that there is an attractive molecular force with a repulsive core. We will look at these more general features first and then return to the details we see in Figure (16).



#### Figure 16

Binding energy. If an atom is in its equilibrium position, we can think of it as sitting at the bottom of its potential energy well. To remove the atom from the molecule, we have to lift it out of the well. Thus the binding energy is the depth of the well. If the atom has a thermal kinetic energy 3/2 kT, and this thermal energy is less than the binding energy, the molecule should stay together.

# Evaporation

Simple features of molecular forces lead to a reasonable understanding of the transition from a liquid to a gaseous state, the process of evaporation. We start with a picture of a liquid as a collection of molecules that all attract each other, can move around past each other, but are nearly incompressible because the repulsive core in the molecular force prevents atoms from being squeezed into each other. The incompressibility of water can be seen from the fact that water in the deepest parts of the ocean, where the pressures are some 800 times atmospheric pressure, is only about 3% denser than the water at the surface.

A molecule in a liquid is free to move around because of its thermal kinetic energy and because there is essentially no net force on it. Although attracted to all of its neighbors, the neighbors surround the molecule as shown in Figure (17), and the net force is zero.

The situation is different for a molecule on the surface as shown in Figure (18). Such a molecule has neighbors only to the sides and below. If we try to lift such a molecule out of the surface, there will be a net force exerted by all of the molecules beneath it, pulling the molecule back in. To extract a molecule from the



## Figure 17

A molecule in the interior of a liquid is attracted by all its neighbors which surround it. As a result the net force is zero and the molecule is free to move about through the liquid. surface requires that you do work against these attractive forces. The amount of work required to extract a molecule from the surface depends upon the type of liquid and the temperature of the liquid, but some energy is required as long as the surface exists.

## Example 2

To estimate the amount of energy required to extract a water molecule from the surface of water, we note that to boil 1 gram of water requires  $2.25 \times 10^{10}$  ergs of energy. Since there are  $3.3 \times 10^{22}$  molecules in 1 gram of water, this represents an energy of  $7 \times 10^{-13}$  ergs per molecule. Some of the energy you supply goes into displacing the air above the water to make room for the steam, but most of it goes into supplying the energy each molecule needs to escape water at  $100^{\circ}$  C (373K).

## Exercise 5

(a) What is the average kinetic energy of a molecule at a temperature of 373K?

(b) Is this enough energy for an average water molecule to escape through the surface of the water?

(c) At what temperature does the thermal kinetic energy equal to the  $7 \times 10^{-13}$  ergs needed to escape?



## Figure 18

A molecule on the surface is attracted to its neighbors beneath it. To pull a molecule out of the surface, you have to overcome these forces. As a result it takes energy to remove a molecule from the liquid. This surface force is often referred to as surface tension. If you worked Exercise 6, you realize that the average molecule, even in boiling water, does not have nearly enough thermal kinetic energy to escape through the surface. Yet even at room temperature water evaporates; even at these lower temperatures some molecules escape through the surface. The reason is that, while 3/2 kT is the average thermal kinetic energy of the molecules, some molecules have more kinetic energy than average, some less. Some have so much more kinetic energy than average that they can escape.

The rate of evaporation depends very much on the distribution of thermal kinetic energies. At a given temperature T what fraction of the molecules have a kinetic energy sufficiently far above average to be able to escape? It turns out that for a substance in thermal equilibrium, there is a precise formula for the distribution of thermal kinetic energies, a formula known as the *Boltzman distribution* which we discuss in Chapter 22. For now we will not go into that much detail. Instead, we will simply recognize that some molecules are hotter than average, some colder than average, and that it is the very hottest ones that have enough energy to escape.

If it is the hot molecules that escape during evaporation, then the cooler ones must be left behind and *evaporation must be a cooling process*. There must, however, be a net loss of molecules from the surface for cooling to occur. As we noted at the beginning of the chapter, the surface of water is a dynamic place where water molecules are continually leaving and returning. A returning water molecule, even if relatively cool when in the air above the water, gains as much kinetic energy when it reenters the water as the hot molecule lost when escaping. Thus reentering molecules become hot when they get back in the water, and thus the returning or condensation of water molecules is a warming process. Whether you get evaporation or condensation depends upon the number of water molecules in the air above the water. If you cover a glass of water with a dish, soon the number of water molecules in the air in the glass builds up to the point that there is a balance between molecules leaving and molecules entering the liquid surface. When this balance is achieved, evaporation ceases and we say that the air above the water is at 100% relative humidity. In order to get cooling from evaporation, the relative humidity of the air must be less than 100%.

The human body uses evaporation for cooling which is effective on a hot, dry day but not on a humid one. When the relative humidity approaches 100% there is no net loss of water molecules and no cooling. Incidentally, you blow on soup to cool it, not necessarily because your breath is cooler than the soup, but because you are replacing the moist air over the soup with drier air so that more evaporation can take place.

# PRESSURE

When you try to compress a liquid, you are trying to shove the atoms into each other which is very difficult to do because of the repulsive core of the molecular force. It is also hard to compress a gas—try blowing air into a soda bottle. But when you compress air you are not squeezing air molecules together, you are not trying to overcome a molecular force at all. The air that you breathe is mostly empty space, the separation of air molecules being about 10 molecular diameters. There is a completely different explanation for why it is difficult to compress a gas, why a gas exerts a pressure that you must overcome to compress it.



## Figure 19a

The air molecules bouncing off the inside surface of the balloon, transfer an outward directed momentum to the rubber. The average momentum per second transferred by these collisions is the average force the molecules exert on a section of the balloon surface. One of the simplest demonstrations of the pressure exerted by a gas is provided by a rubber balloon. When you blow up a balloon, the rubber of the balloon is trying to compress the gas, force it down to a smaller volume. The molecules of the gas exert an outward force on the rubber, preventing it from collapsing. This outward force is caused by the collisions of the gas molecules with the rubber, as illustrated in Figure (19). Whenever an air molecule strikes and bounces off of the rubber, there is a net transfer of outward directed linear momentum to the rubber. Since the collisions are occurring continually, there is a continual transfer of momentum to the rubber, which, according to Newton's second law  $\vec{F} = d\vec{p}/dt$ , means that the molecules exert a force on the rubber. On the average, the direction of momentum transfer is outward, perpendicular to the surface of the rubber. Thus the force exerted by the gas molecules is also perpendicular to the surface.

The total force exerted on some part of the balloon depends upon the area of the balloon we are talking about. We can simplify the discussion by talking about the *force on a unit area* of the balloon surface, and give that force the special name *pressure*. We know that force is a vector quantity, but the force that a gas exerts on a surface is always directed perpendicular to the surface, no matter what orientation the surface has.

## Figure 19b

Balloon placed on liquid nitrogen. If you cool the air molecules inside the balloon, they do not strike the rubber as hard, exert less of a force, and the balloon collapses. In the final picture, there is only a puddle of liquid air inside.







Thus we can let the word "pressure" stand for the magnitude of the force on a unit area, and determine the direction of the force from the orientation of the surface.

$$\frac{\text{pressure}}{\text{of a gas}} = \frac{\text{magnitude of the force}}{\text{exerted by the gas on a}}$$
(3)

In the case of a gas inside a balloon, it is clear that the behavior of the gas molecules is more or less the same throughout the balloon. As a result, the pressure should be essentially the same on all surface areas of the balloon. Instead of using the word "pressure" to merely describe the force on surfaces of the balloon, we can say that *the pressure is in the gas itself*. Once you know the pressure of the gas, you can then calculate the force the gas exerts on some particular surface by multiplying the pressure times the area of the surface, and noting that the force is directed perpendicular to the surface.

The pressure of the gas, the force the gas molecules exert on a surface, depends upon how fast the molecules are moving when they hit the surface. The faster their average speed, the greater the force and pressure. Since the motion of the molecules is thermal motion, whose average kinetic energy is 3/2 kT, the average speed and pressure must increase with temperature. If you heat the balloon, the gas pressure increases and the balloon expands. If you cool it, it contracts.

An excellent demonstration of the dependence of air pressure on temperature is to place a balloon in a bucket of liquid nitrogen, as shown in Figure (19b). A common Styrofoam ice bucket makes an excellent container for liquid nitrogen for this demonstration. When you place the balloon on the surface of the liquid nitrogen, the balloon sits there for a while, and then begins to shrink as the air molecules cool down. The shrinking continues until the balloon collapses and all you have inside is a puddle of liquid air. Now any further contraction of the balloon would require squeezing the air molecules themselves together which is opposed by the repulsive core of the molecular forces. In a sense, the collapse of the cooling gas in the balloon was halted by molecular forces. If you take the balloon out of the liquid nitrogen, the air inside warms up and the balloon expands again to its original volume. Remarkably, a typical balloon can undergo this cycle a number of times without breaking.

## **Stellar Evolution**

The balloon has features in common with our sun. The sun is not the solid object it appears. Abetter model is that the sun is a bag of gas, somewhat like a balloon, but with the constraining force of the rubber replaced by gravity. The sun looks like it has a distinct surface, but that is an illusion created by the fact that the gas molecules in the sun, which are mostly hydrogen, become ionized and opaque at temperatures in excess of 3000 Kelvins. As you go down into the sun, the temperature of the gas increases. The distance at which it reaches 3000 Kelvins, the gas changes from transparent to opaque and that is what we see as the surface of the sun.

For the past five billion years, the sun has gotten energy from the conversion of hydrogen nuclei to helium nuclei. And there is another five billion years worth of hydrogen left before the sun runs out of fuel. At that point the sun will do something spectacular. It will expand so that the earth will be orbiting near the sun's surface. (We discuss this process in more detail in Chapter 20.) But eventually the sun will settle down and begin to cool off.

As the sun cools, it will contract very much like the balloon in our demonstration. And like the balloon, the collapse will be halted when the atoms are so close together that the repulsive core of the molecular forces prevents further contraction. At that point the sun will have become what is called a **white dwarf**, an object about the size of the earth slowly cooling until it becomes a dark ember.

If the sun were just a bit bigger, about 1.4 times its current mass, the gravitational collapse would not be halted by the molecular forces between atoms. That is the mass at which gravity is strong enough to crush the atoms together, to overcome the atomic repulsive cores, with the result that you end up with a neutron star. That is also a topic we discuss in Chapter 20.

# THE IDEAL GAS LAW

We have discussed the picture of how the repeated collisions of the gas molecules inside a balloon exert an outward force on the rubber, and how, if we raise the temperature, the molecules travel faster, strike harder, and exert a greater force. We will now, with a fairly simple derivation, obtain an explicit relation between the temperature of the molecules and the force they exert, a relationship known as the *ideal gas law*.

There are many ways to derive the ideal gas law, depending upon what assumptions you are willing to make about averaging over molecular speeds. The less you are willing to assume, the harder the derivation is. But there is one rather surprising feature of all the derivations. They all give the same correct answer. The usual procedure in textbooks is to make the derivation as complex as students will tolerate, apologize for or hide the approximations, and announce that the answer is correct. What we will do is present the simplest derivation we can find that gives the right answer. When an argument looks too simple to be true, but gives the right answer, that means that you may have extracted an important basic feature from a complex situation.

## **The Ideal Piston**

While a balloon is a very practical container for gas, the curved surfaces make it a bit difficult to use for theoretical analysis. Instead we will, more or less as a thought experiment, use an idealized device called the friction-less piston. Figure (20) is a diagram of a frictionless piston in a cylinder of cross-sectional area A. In the cylinder is a gas -- like air at room temperature. The gas molecules are bouncing around, colliding with the walls of the cylinder and the face of the piston. Because of the collisions, the gas molecules exert a force on the piston, and because the piston is frictionless, we must exert an oppositely directed force  $\vec{F}$ , as shown, to keep



Figure 20

Frictionless cylinder in a piston. Picture the force that the gas molecules exert on the piston as being counterbalanced by an external force  $\vec{F}$  as shown.

the cylinder from expanding. We are assuming that there is no gas behind the piston, so that only the force  $\vec{F}$  keeps the gas from expanding.

We know that the force F exerted by the gas increases with temperature because the balloon expanded when we heated it. We also know that the average thermal kinetic energy of the gas molecules increases as 3/2 kT as the temperature rises. What we wish to do now is relate these observations to obtain a formula for how the force F depends upon the temperature T.

To relate F to T we start with the simplest possible model of the gas in the cylinder, namely a gas consisting of one molecule, bouncing back and forth at a speed v, as shown in Figure (21). Each time the molecule strikes and bounces off the piston, its momentum changes by 2mv. Since linear momentum is conserved during the collision, the piston picks up an outward, x-directed, linear momentum of magnitude 2mv as a result of the collision. (Remember problem 7-5, where two skaters on frictionless ice were tossing a ball back and forth. When one of the skaters caught the ball, she picked up the ball's momentum mv. When she threw the ball back, she recoiled, picking up an additional mv of momentum. As a result she picked up 2mv of momentum with each catch and toss.)

If we designate by  $\Delta p$  the magnitude of the x-directed linear momentum that the piston gains from each collision we have

$$\begin{array}{ll} momentum \\ transferred \\ per collision \end{array} \Delta p = 2mv \tag{4}$$

The momentum  $\Delta p = 2mv$  is transferred to the piston each time the molecule comes back and strikes the piston. If  $\Delta t$  is the time between collisions, then the amount of momentum transferred per second is  $\Delta p/\Delta t$ .



**Figure 21** Analysis of a one molecule gas.

Using Newton's second law, in the form  $\vec{F} = d\vec{p}/dt$ , we see that this rate of transfer of linear momentum  $\Delta p/\Delta t$  is just the average force F that the molecule is exerting on the piston

average force		average rate	
exerted by	$E - \Delta p$	of momentum	
molecule	$\Gamma = \overline{\Delta t}$	transferred	(5)
on piston		to the piston	

To calculate the time  $\Delta t$  between collisions, note that if the distance from the end of the cylinder to the piston is l, and the molecule is traveling at a speed v, it covers the distance down and back, 2l, in a time

$$\Delta t = \frac{2\ell \text{ cm}}{\text{v cm/sec}} = \frac{2\ell}{\text{v}} \text{sec}$$
(6)

With Equation (4) for  $\Delta p$  and (6) for  $\Delta t$  in (5), we get, for the average force F exerted by this one molecule of gas

$$F = \frac{\Delta p}{\Delta t} = \frac{2mv}{(2\ell/v)} = \frac{mv^2}{\ell}$$
(7)

Note the appearance of  $mv^2$  in Equation 7. We are already beginning to see a connection between the molecule's kinetic energy and the force it exerts.

If you could actually set up a one molecule, one dimensional, gas like that shown in Figure (21), Equation 7 would accurately describe the average force of that gas on the piston. No approximations have been made yet. The approximations enter when we go to a gas of N molecules, moving in three dimensions, at various speeds. The simplest, most outrageous approximation we can make is that all of the molecules have the same speed v, and that 1/3 of them are bouncing back and forth in the x-direction, 1/3 in and out in the y-direction, and 1/3 up and down in the z-direction. Such a gas with N/3 molecules bouncing back and forth, would exert a force N/3 times as great as our one molecule gas in Figure (21)

$$F\begin{bmatrix} N \text{ molecule} \\ gas \end{bmatrix} = \frac{N}{3} F\begin{bmatrix} 1 \text{ molecule} \\ gas \end{bmatrix}$$
$$= \frac{N}{3} \left( \frac{mv^2}{\ell} \right)$$
$$= \frac{2}{3\ell} \times N \times \left( \frac{1}{2} mv^2 \right)$$
(8)

Since, in this approximation, all the molecules have the same speed v, then the factor  $1/2 \text{ mv}^2$  in Equation 8 must be the average thermal kinetic energy of the molecules. Replacing  $1/2 \text{ mv}^2$  by 3/2 kT gives

$$F\begin{bmatrix}N \text{ molecule}\\gas\end{bmatrix} = \frac{2}{3\ell} N \times \frac{3}{2} kT$$
(9)

Now you see why the factor of 3/2 was inserted into the formula 3/2 kT for the average thermal kinetic energy. The 3/2 cancels the 2/3 that appeared in Equation 9, and we get

$$\mathbf{F}\begin{bmatrix}N \text{ molecule}\\gas\end{bmatrix} = \frac{\mathbf{N}}{\ell} \mathbf{k} \mathbf{T}$$
(10)

We are almost finished. Equation 10, despite our approximations, is the correct formula for the force exerted by a gas of N molecules at a temperature T. The only problem with Equation 10 is the explicit dependence on the length  $\ell$  of the cylinder. We can remove this explicit dependence by expressing the force on the cylinder in terms of the pressure P of the gas.

In our earlier discussion, we said that the pressure of a gas inside a balloon was equal to the force per unit area exerted by the gas on the surface of the balloon (Equation 3). If we have a gas at pressure P in a cylinder of cross-sectional area A, as shown in Figures (20) and (21), then the force exerted on the piston, whose area is A, must be

$$\mathbf{F} = \mathbf{P}\mathbf{A} \qquad \begin{array}{c} pressure\\ times area \end{array} \tag{11}$$

Substituting Equation 11 for F in Equation 10, and multiplying through by the cylinder length l, gives

$$P(A\ell) = N kT$$
(12)

The final step is to note that  $(A\ell)$ , the area of the cylinder times its length, is the volume V of the cylinder. Thus we get

$$PV = NkT \qquad ideal \ gas \ law \qquad (13)$$

Equation 13 is known as the *ideal gas law*. Despite the approximations we used to derive it, it is accurate as long as the particles in the gas are separated enough that one can neglect the molecular forces between particles. To express this another way, any gas that obeys Equation 13 is known as an *ideal gas*. To a very high degree of accuracy, the air around us behaves as an ideal gas.

# **Ideal Gas Thermometer**

The ideal gas law, PV = NkT, incorporates such laws as Boyle's law and Charle's law which you may have encountered in an introductory chemistry course. One can construct numerous examples and homework problems applying this law. What we will do for our first application is to describe the ideal gas thermometer which we will use, at least for now, as our experimental definition of temperature.

An example of an ideal gas thermometer is shown in Figure (22). The glass tube and the plug of mercury come about as close as you can get to a cylinder with a frictionless piston. You can make one of these devices by sealing a glass tube at the bottom, pouring some mercury in, evacuating most of the air above the mercury, and sealing the top of the tube. If the tube is fairly small, when you turn the tube over, the mercury



## Figure 22

Ideal gas thermometer. If we heat the gas beneath the mercury plug, the gas expands raising the plug. If we have an ideal gas, then the height of the plug depends only on the temperature of the gas and not the kind of gas we used. plug will slide down until it sits on the remaining air. There will be a vacuum above the plug. How high the plug rides depends upon the length  $\ell$  of the mercury plug and how much gas you left in the tube before sealing it.

We can use the ideal gas law to predict how the height of the plug varies with the temperature of the gas in the tube. When we do this, we obtain a messy looking formula with factors like the density  $\rho$  of the mercury, the number N of air molecules in the tube, the area A of the tube, the acceleration g due to gravity and Boltzman's constant k. But when we take another look at the result we see that most of the factors are constants, and the height h of the air column turns out to be strictly proportional to the temperature T of the gas in the tube. Let us see how this all works out.

Rewriting the ideal gas law as an equation for the temperature T of the gas molecules, we get

$$T = \frac{PV}{Nk} = \frac{PhA}{Nk}$$
(14)

where V = hA is the volume occupied by the air which is in a column of height h and area A. The mercury plug of length lriding on top of the gas, exerts a gravitational force mg on the gas, where the mass m of the mercury is equal to the mercury's density  $\rho$  times its volume lA. Thus

weight of mercury column 
$$\rangle$$
 = mg =  $\rho \ell Ag$  (15)

The force mg is the total force exerted by the mercury column on the air. The force per unit area, which must equal the pressure of the gas if the plug is balanced on the gas is

pressure of gas  
beneath a plug  
of mercury  
of length 
$$\ell$$
  $P = \frac{mg}{A} = \frac{\rho \ell A g}{A} = \rho \ell g$  (16)

Equation 16 will turn out to be useful in other experiments, for it tells us how to measure the pressure of a gas in terms of the height  $\ell$  of a mercury plug that the gas can support.

$$T = \frac{PhA}{Nk} = \left(\frac{\rho \ell gA}{Nk}\right)h = \chi h$$
(17)

where  $\chi = \rho \ell g A/Nk$  is a collection of constants. The basic result is that the gas temperature T is strictly proportional to the height h of the air column.

To use an ideal gas thermometer, we do not need to evaluate the constants in the formula for  $\chi$ . Instead immerse the thermometer in ice water and mark the bottom of the plug 0°C. Then put the thermometer in boiling water and mark that 100°C. Mark off the distance between 0°C and 100°C in 100 equally spaced intervals and you have a centigrade thermometer.

The fascinating feature of an ideal gas thermometer is that you can quickly determine the temperature at which the gas volume should go to zero, the temperature we have called *absolute zero*. On a sheet of graph paper, mark off a temperature scale on the bottom that runs backwards from  $100^{\circ}$  C to  $0^{\circ}$  C and goes on out quite away into negative temperatures. On the vertical axis plot the height h of the air column. For this plot,

Height of

100°C

air column

you have only 2 experimental points, the height at 0° C and at 100° C. Connect these two points by a straight line (that is what the formula  $T = \chi h$  says you should do), and you find that h goes to zero at a temperature of  $-273^{\circ}$  C. That is all there is to it!

From our discussion of molecular forces, you can see that any ideal gas thermometer you actually build has to fail before you get to absolute zero. At some point as you cool the air in the thermometer, you end up with a puddle of liquid air as we did in the balloon demonstration. Even before the air becomes liquid, the spacing between the air molecules is reduced to the point where the molecular forces between air molecules becomes important. The attractive molecular forces reduce the pressure of the gas, the gas no longer obeys the ideal gas law, and we cannot believe the readings of the thermometer. This problem can be put off by using helium gas that remains a gas down to a temperature of 4 kelvins, but that's the limit. To work at temperatures closer to absolute zero you need a different experimental definition of temperature, like the thermodynamic definition we discuss in Chapter 18.

#### Figure 23

Absolute zero. If you plot the height of the air column in an ideal gas thermometer as a function of temperature, drawing a straight line between the two known data points at 100 °C (boiling water) and 0 °C (melting ice), and continue the line down to zero height, the intersection is at -273 °C. This represents an absolute low value for temperature, as defined by the ideal gas thermometer.



## The Mercury Barometer and Pressure Measurements

Columns of mercury are useful not only for making thermometers, but also for making devices to measure pressure. We will begin with a discussion of the mercury barometer whose construction is shown in Figure (24).

As shown in Figure (24a), start with a u shaped glass tube about a meter long, sealed at one end, and work mercury into it until the sealed section is nearly full. Then invert the tube as shown in Figure (24b). The mercury in the sealed section will slide down, leaving a vacuum behind it, until the difference in the heights of the mercury columns is about 76 cm as shown.

The height difference of the two columns tells us the pressure of the atmosphere. To see why, conceptually break the mercury column up into two parts as shown in Figure (24c). The bottom part is the loop of mercury that goes from point 1 (at the open end of the mercury) to point 2 (at the equal height in the closed section). The upper part, goes from point (2) up to the vacuum, a section whose height we designate by the letter h. This column sits over the bottom loop and exerts a downward force equal to the weight mg of a column of mercury of height h.

The mercury in the bottom section between points (1) and (2), is completely free to move up one side or the other. Since it does not move, the weight mg of the mercury column pushing down on the left side at point (2) must be balanced by the force of the atmosphere pushing down at the open end, point (1). As indicated

in Figure (25), the molecules of the air are colliding with the surface of the mercury, exerting a force in the same way that the air molecules in a balloon push out on the rubber. If the air molecules are at a pressure  $P_a$  (pressure of the atmosphere) and the glass tube has an area A, the force exerted by the air is the pressure times the area.

force exerted  
by atmosphere  
on air column 
$$= P_a A$$
 (18)

The weight mg of the mercury column pushing down at point (2) is

$$mg = \rho(Ah)g \tag{19}$$

where  $\rho$  is the density of mercury and Ah is the volume of mercury in the column of area A and height h.



#### Figure 25

The weight of the mercury column above point (2) must be balanced by the force exerted by the atmosphere at point (1).

#### Figure 24

Construction of a mercury barometer. When you turn the tube over, going from (a) to (b), the mercury slides down the sealed leg, leaving a vacuum behind. The difference in heights h of the two columns (c) is a measure of atmospheric pressure.



Equating the forces on the two sides of the mercury in the bottom section gives

$$P_{a}A = mg = \rho hAg$$

$$P_{a}A = \rho gh$$
(20)

The result is that the atmospheric pressure is proportional to the height difference h in the two columns of mercury. As we have mentioned, the dimensions of pressure in CGS units is dynes per square centimeter, while in MKS units it is Newton's per square meter, a set of dimensions given the name *pascal*. Neither set of units is particularly convenient. Using  $\rho = 13.6$  gm/cm<sup>3</sup> for the density of mercury, and using the value h = 76 cm for an average value for the height of the mercury column, we get

$$P_{a}A = 13.6 \frac{gm}{cm^{3}} \times 980 \frac{cm}{sec^{2}} \times 76 \text{ cm}$$
$$= 1.01 \times 10^{6} \frac{gm \text{ cm/sec}^{2}}{cm^{2}}$$

$$P_a A = 1.01 \times 10^6 \text{ dynes / cm}^2$$
 (21a)

Converting to MKS units, where one newton equals  $10^5$  dynes and  $1 \text{ m}^2 = 10^4 \text{ cm}^2$  we have

$$P_{a}A = 1.01 \times 10^{6} \frac{\text{dynes}}{\text{cm}^{2}} \times \frac{10^{4} \text{ cm}^{2}/\text{m}^{2}}{10^{5} \text{ dynes/newton}}$$
$$= 1.01 \times 10^{5} \frac{\text{newtons}}{\text{m}^{2}}$$
$$= 1.01 \times 10^{5} \text{ pascals}$$
(21b)

Just as it was inconvenient to measure temperature in ergs, it is rather inconvenient for the weatherman to announce today's barometric pressure in either dynes per square centimeter or pascals. Numbers in the range of  $10^6$  or  $10^5$  do not go over well with the listening audience.

What is much easier is to simply express the pressure in terms of the height of the mercury column that the atmospheric pressure will support. "A low pressure system moved over the area today, and the barometer reading dropped to 752 millimeters of mercury" sounds much better than saying that the pressure dropped to  $1.0023 \times 10^5$  pascals.

The millimeter of mercury, as a unit of pressure, has been given the name *torr*, one more name inflicted upon students by those who decide what the standard names shall be. For comparison's sake, we can express atmospheric pressure as

$$P_{a} = 1.01 \times 10^{6} \text{ dynes/cm}^{2}$$
  
= 1.01 × 10<sup>5</sup> pascals  
= 101 kilo pascals  
= 76 cm Hg (22)  
= 760 mm Hg  
= 760 torr  
= 14.7 lbs/in<sup>2</sup>

At different times you may encounter any of these units.

When you are working with vacuum pumps and vacuum gauges, even the torr, 1 millimeter of mercury, is too large a unit to be convenient. Many gauges are calibrated in *microns*, which is the pressure exerted by one micron or one millionth of a meter of mercury.

1 "micron" = 
$$10^{-6}$$
 meters Hg  
=  $10^{-4}$  cm Hg  
=  $10^{-3}$  mm Hg (23)  
=  $10^{-3}$  torr

In the electron gun experiments we discuss in Chapter 28, the glass tube containing the electron beam is evacuated to a pressure of around one micron. Current technology allows you to work with much better vacuums in the range of  $10^{-6}$  to  $10^{-7}$  microns. Such vacuums are needed to maintain clean surfaces when studying the atomic structure of surfaces or creating complex electronic chips.

## **Exercise 6**

What is the pressure P in dynes/cm<sup>2</sup> and pascals, inside an apparatus where the pressure gauge reads

a) 1 cm Hg

- b) 1 micron
- c) 1 kilo pascal
- d) 1 torr
- e) 10<sup>-9</sup> torr
- f) 50 microns
- g) 30 lbs/in<sup>2</sup> (U.S. tire pressure gauge)
- h) 200 kilo pascals (European tire pressure gauge)

# AVOGADRO'S LAW

Speaking of inconvenient units like measuring temperature in ergs or pressure in dynes per  $cm^2$ , we have something particularly inconvenient in the form of the ideal gas law PV = NkT. To use this equation, we have to know the number N of the molecules in the gas. To actually count the molecules is essentially an impossible requirement.

Instead of counting individual molecules, we can lump them in large units, and count the number of units. The standard unit for counting molecules is the *mole*. If you have a mole of molecules or any other object, you have  $6 \times 10^{23}$  of them.

1 mole of objects = 
$$6 \times 10^{23}$$
 objects (23)

The idea of a mole is that it is a convenient counting device for handling large numbers. You might, for example, hear an astronomer say that there is about a mole of stars in the visible universe. By that the astronomer would mean that he thinks that the visible universe contains about  $6 \times 10^{23}$  stars. (That estimate may not be too many orders of magnitude off.) As another example, it would take about a mole of baseballs to fill the volume of the earth with baseballs.

The number  $6 \times 10^{23}$  (more accurately  $6.02 \times 10^{23}$ ), which is known as *Avogadro's number* or constant, is essentially the number of hydrogen atoms in one gram of hydrogen. Since hydrogen atoms and protons have essentially the same mass, a mole of protons also has a mass of 1 gram.

When you have a bottle of hydrogen gas, the hydrogen atoms combine in pairs to form hydrogen molecules. Thus a mole of hydrogen molecules has a mass of 2 grams. An oxygen atom is 16 times as massive as a hydrogen atom, thus a mole of oxygen molecules, with 2 atoms in each molecule, has a mass of  $2 \times 16 = 32$  grams. The mass of a mole of a given kind of molecule is usually called the *molecular weight*, but more properly the *molecular mass*, of that kind of molecule.

(The integer numbers that appear in the mass of atoms arises from the fact that protons and neutrons which make up the atomic nucleus have about the same mass, and the mass of the electrons is much much smaller. Most oxygen atoms for example have a nucleus with 8 protons and 8 neutrons, and that is why an oxygen atom is 16 times as massive as a hydrogen atom.)

We will use the symbol  $\mathbf{N}_{\mathbf{A}}$  to designate Avogadro's number

$$N_A = 6 \times 10^{23} \frac{\text{particles}}{\text{mole}} \quad \stackrel{Avogadros}{number}$$
(24)

We can now rewrite the ideal gas law in the form

$$PV = NkT = \frac{N}{N_A} \times (kN_A)T$$
(25)

We do this because  $N/N_A$  is the number of moles of the substance rather than the number of molecules. Designating this by the symbol n, we have

$$n \equiv \frac{N \text{ molecules}}{N_A \text{ molecules/mole}} = \frac{N}{N_A} \text{ moles}$$
 (26)

In addition the product of Boltzman's constant k times Avogadro's number  $N_A$  is called the gas constant R

$$R \equiv kN_{A} \quad (gas constant)$$
  
=  $1.38 \times 10^{-23} \frac{joules}{kelvin} \times \frac{6.02 \times 10^{23}}{mole}$  (27)  
$$R = 8.31 \frac{joules}{mole kelvin} \qquad MKS units$$

In this case the MKS units are much more convenient. The gas constant R in the CGS system is  $10^{-7}$  times smaller.

Expressing the ideal gas law in terms of the number of moles n and the gas constant R gives

$$PV = nRT$$
(28)

which is the alternate form of the ideal gas law. If your perspective is from an atomic point of view, you would use the form PV = NkT. But if you were a chemist and had to actually measure the quantities involved, you would use the form PV = nRT.

Our first example of the use of Equation 28, will be to determine the volume of one mole of molecules at  $0^{\circ}$  C (273 K) and atmospheric pressure ( $1.01 \times 10^{5}$  pascals). We have

$$PV = nRT$$
  
1.01 \* 10<sup>5</sup> newtons × V  
= 1 mole × 8.31 joules × 273 K

First let us check the dimensions. The moles and the kelvins cancel on the right side, and we get

V ~ joules 
$$\times \frac{m^2}{newton} = kg \frac{m^2}{sec^2} \times \frac{m^2}{kg m/sec^2}$$
  
= meter<sup>3</sup>

The numerical value is

$$V = \frac{8.31 \times 273}{1.01 \times 10^5} \text{ m}^3$$
$$= 22.4 \times 10^{-3} \text{ m}^3$$

Noting that  $10^{-3}$  m<sup>3</sup> is one liter, we get

$$V = 22.4 \text{ liters}$$

$$volume of 1 mole 
of any gas at 0° C 
and atmospheric 
pressure
$$(29)$$$$

The important point is that a mole of any kind of gas has a volume of 22.4 liters at the standard conditions of 0° C and atmospheric pressure. (One often uses the notation *STP* for this *standard temperature and pressure*.) At STP, 22.4 liters of hydrogen have a mass of 2 grams, nitrogen 28 grams, and oxygen 32 grams.

## **Exercise 7**

A helium nucleus contains 2 protons and 2 neutrons. The mass of 22.4 liters of helium gas at STP is 4 grams. What does that say about the molecular force between helium atoms?

The calculation of the volume of a mole of a gas emphasizes an important point about the behavior of an ideal gas. Namely, if we have equal volumes of gas at the same temperature and pressure, the volumes will contain the *same number of molecules*. This was first suggested by the Italian scientist Amedeo Avogadro (1776--1856), and is known as *Avogadro's law*.

# HEAT CAPACITY

From our discussion of temperature and other processes from an atomic and molecular point of view, it is obvious that the way to raise the temperature of an object is to add energy. At higher temperatures the average thermal kinetic energy of the molecules increases, and that energy must come from somewhere. Historically the relationship between heat and energy was not so clear. As late as 1798, 71 years after Newton's death, there were accepted theories that treated heat as a substance called *caloric* that flowed from hot substances to cooler ones.

It was Benjamin Thomson, later known as Count Rumford, who proposed that heat was, in fact, a form of energy. Rumford was boring cannons for Prince Maximilian of Bavaria, and was quite aware that when the drills were dull, the cannons became hot. Thomson proposed that the mechanical work he put into turning the drills was converted to heat energy that raised the temperature of the cannons. Forty years later, Joule accurately measured the amount of work required to raise the temperature of various substances.

Traditionally heat energy was defined as the amount of heat required to raise the temperature of one gram of water one degree centigrade. This unit of heat is called the *calorie*. In terms of mechanical energy, the conversion factor is

$$l \text{ calorie} = 4.186 \text{ joules} \tag{30}$$

This is the relationship between mechanical work and heat that Joule studied.

# Exercise 8

A 1 kilogram mass is dropped into a bucket containing 1 liter ( $10^3$  cm<sup>3</sup>) of water. Assume that all of the kinetic energy of the mass ends up as heat energy, raising the temperature of the water.

(a) From what height would you have to drop the mass to raise the temperature one degree centigrade?

(b) (More realistic question.) How much would the temperature rise if you dropped the mass from a height of one meter?

# **Specific Heat**

The amount of heat energy required to raise the temperature of a unit mass of a substance one degree is called the *specific heat capacity* or specific heat of the substance. For example, since it requires one calorie to heat one gram of water one degree centigrade, we can say that the specific heat of water is 1 calorie/gm °C, or 4.186 joules/gm °C.

# **Molar Heat Capacity**

If, instead of measuring the heat capacity of a unit mass, we measure the heat capacity of a mole of a substance, we call the result the *molar heat capacity*. For example a water molecule  $H_2O$  with 2 hydrogen and 1 oxygen atom is 18 times as massive as a hydrogen atom. Its molecular weight is 18, and thus a mole of water has a mass of 18 grams. As a result it takes 18 calories to raise the temperature of a mole of water 1 degree centigrade, and thus the molar heat capacity of water is 18 calories/ mole °C or  $18 \times 4.186 = 75.3$  joules/mole °C. For the units, instead of degrees centigrade, we can use kelvins, which are the same size. Thus we can write

$$\begin{array}{c} \text{molar specific} \\ \text{heat of water} \end{array} \right\} = 75.3 \frac{\text{joules}}{\text{mole K}}$$
(31)

as an example of a molar specific heat.

Predicting the specific heat of a substance, even with an understanding of the atomic and molecular processes involved, turned out to be a much more difficult subject than expected. The first time a failure of Newtonian mechanics was detected was during the efforts to predict the specific heats of various gases. This failure was due to quantum mechanics being necessary to fully understand what happened to the added heat energy.

There is one example, however, where the simple picture of atoms we have been discussing gives the correct answer. That is for the specific heat of helium gas. We will discuss that example here, and leave all other discussions of specific heat to Chapter 20, an entire chapter devoted to the subject.

## Molar Specific Heat of Helium Gas

A gas of helium atoms is about the simplest substance you can picture. Since helium does not form molecules, the gas simply consists of individual atoms moving around and bouncing off of each other. If the temperature of the gas is T, then the average thermal kinetic energy of the atoms is 3/2 kT.

If you have a mole of helium gas at a temperature T, then the thermal energy of the atoms should be the average energy of 1 atom, 3/2 kT, times the number N<sub>A</sub> atoms in a mole. Thus we easily estimate that the thermal energy E<sub>He</sub> of a mole of helium atoms is

$$E_{\text{He}} = N_{\text{A}} \times \left(\frac{3}{2} \text{ kT}\right)$$
$$= \frac{3}{2} \left(N_{\text{A}} \text{k}\right) \text{T}$$

Using the fact that  $N_A k = R$ , the gas constant, we get

$$E_{\text{He}} = \frac{3}{2} \operatorname{RT} \qquad \begin{array}{c} \text{thermal energy} \\ \text{of a mole of} \\ \text{helium atoms} \end{array} \tag{32}$$

If we raise the temperature one degree, from T to (T + 1), the thermal energy goes from 3/2 RT to 3/2 R(T + 1), an increase of 3/2 R. Thus the molar specific heat, which we will call C<sub>V</sub>, is

$$C_{V} = \frac{3}{2}R = \frac{3}{2} \times 8.31 \frac{\text{joules}}{\text{mole } K}$$

$$C_{V} = 12.5 \frac{\text{joules}}{\text{mole } K} = \frac{3}{2}R$$
(33)

As we mentioned, we get the right answer. Equation 33 is in agreement with experiment.

The subscript V on the symbol  $C_V$  is there to remind us to measure the specific heat at constant volume. If you add heat to a gas, and at the same time allow the gas to expand, some of the energy goes into the work required to expand the volume, pushing the surrounding gas aside. This is a complication that we will discuss in Chapter 18. For now we will leave the subscript V on  $C_V$  to remind us not to let the volume increase.

## **Other Gases**

It took almost no effort to correctly predict the specific heat of helium gas. What complications do we face when we try to predict the specific heat of other gases?

The problem is that other gases form molecules. From one point of view the molecules themselves are the gas particles, so that their average thermal kinetic energy must be 3/2 kT just like the helium atoms. So far so good, but molecules have an internal structure. An oxygen molecule, for example, consists of two oxygen atoms held together by the molecular force we discussed back in Figures (14, 15, and 16). As we saw in Figure (15), we can fairly accurately picture the molecule as two atoms held together by a spring force as shown here in Figure (26).

If an oxygen molecule collides with another molecule in the gas, one would expect that the molecule would start to vibrate, and perhaps rotate. This vibration and rotation represent forms of internal motion of the molecule that are quite distinct from the motion of the molecule as a whole, distinct from what we would call the center of mass motion.

If the center of mass motion has an average thermal kinetic energy 3/2 kT just like helium atoms, but the molecules can have internal motions and internal energy, you would expect that it would require more energy to heat a mole of oxygen than a mole of helium. For with the oxygen you not only have to supply the kinetic energy of the center of mass motion, but also the internal energy of the molecules. And that is correct. The molar specific heat of oxygen is 20.8 joules/(mole K), as compared to 12.5 joules/(mole K) for helium.

However it is when we try to calculate how much the internal energy of the molecules contribute to the specific heat, we run into trouble. As far back as 1858, James Clerk Maxwell, who was working on these calculations, repeatedly failed, and suspected that the failure was due to a problem with Newtonian mechanics.



**Figure 26** *Model of an oxygen molecule.* 

# **EQUIPARTITION OF ENERGY**

The success of our calculation of the specific heat of helium gas lies in the fact that in our discussion of thermal processes, the helium atom can be treated as a rigid, undeformable sphere. Striking a helium atom is more analogous to striking a golf ball than hitting whiffle ball. When you hit a golf ball, most of the energy of the impact goes into the kinetic energy of the ball, and the motion of the ball is quite predictable. Hit a whiffle ball and most of the energy goes into mushing the ball; predicting where the whiffle ball will go is difficult.

All gas atoms except helium form molecules. While the individual atoms can usually be treated as hard spheres, for thermal calculations, the molecule as a whole is generally not rigid. Strike a molecule and some of the energy goes into center of mass motion of the molecule, but some goes into internal motions of the individual atoms. It seems that it would be rather hard to say much about the energy of an object that is vibrating, rotating, and flying through space.

However, if you have a gas of molecules in thermal equilibrium, the laws of Newtonian mechanics combined with the mathematical laws of probability, make a surprisingly simple prediction of where the energy goes. This prediction is called the *equipartition of energy theorem* which we will now describe.

As a background for the concepts involved in the equipartition of energy theorem, let us go back to the normal modes experiment of Chapter 16 where we had two air carts on an air track, connected by springs as shown in Figure (16-3). We found that the air carts had two distinct kinds of motion. There was the high frequency mode of motion where the two air carts oscillated against each other, moving together and then apart in a sinusoidal motion. Then there was the low frequency sloshing mode where the carts went back and forth along the track more or less together.

When we started the carts moving in a random way, recorded the motion, and did a Fourier analysis, we found that the apparently complex motion was merely a combination of the two simple sinusoidal modes of motion, the so-called *normal modes*. The carts were not free to move in an arbitrary way, their motion had to either be all of the vibrational mode, or all of the sloshing mode, or some combination of the two. The only thing that was arbitrary about the motion of the carts was how much of each of the two normal modes was present.

In our earlier discussion of center of mass motion in Chapter 11, we considered the example of two air carts, joined to each other by a spring, but free to move down the track as shown in Figure (11-9). We saw that when we gave one of the carts a shove, the center of mass of the two carts moved at a uniform speed down the track, while the carts themselves oscillated about the center of mass. In this example we again have two normal modes of motion. One is the motion of the center of mass, and the other is the oscillation about the center of mass. If we shove the carts just right we can have pure center of mass motion. Or we can have the carts oscillate with no center of mass motion. Or we can have some combination of the two kinds of motion.

These examples begin to show a pattern. If you have two masses connected by springs, that are constrained to move on a one dimensional track, the objects will have precisely two normal modes of motion. What the actual modes are depends upon the way the springs are connected. When the springs were connected to the ends of the air track, there was no center of mass motion, but we had two vibrational modes. When the carts were free to move down the track, we had the mode representing center of mass motion, but only one vibrational mode.

Another term often used to describe the way these carts are moving is the expression *degrees of freedom*. Two carts moving in one dimension have 2 degrees of freedom of motion. The degrees of freedom are the center of mass motion and the vibrational mode shown in Figure (11-9), or the two vibrational modes that we get from the setup in Figure (16-3).

/////**/**/////

Figure 11-9 Oscillating carts

If one connects three air carts with springs and starts them moving, the resulting motion can be quite complex. But if you record the motion, select one cycle of the repeating pattern and do a Fourier analysis, you get the simple result that there are three normal modes, as seen in the results from a student project shown in Figure (16-30). What is not so easy is to find out what the individual normal modes are. And for our discussion now, it is not important to find them. The important result is that if you have three carts connected in some way by springs, and they are constrained to move in one dimension there will be three normal modes or degrees of freedom. (If you have no springs, you still have 3 degrees of freedom, namely the center of mass motion of each of the 3 carts.

The counting of normal modes or degrees of freedom generalizes to more than one dimension. If you have one particle moving in three dimensions, it has three degrees of freedom, one for center of mass motion in the x direction, one for center of mass motion in the y direction and one for center of mass motion in the z direction.

If we have two particles in 3 dimensions, there are 6 degrees of freedom. If they are independent particles, then each has three degrees of freedom of motion of the center of mass. If they are connected by a spring, which is a good model of a diatomic molecule, there are still 6 degrees of freedom but we count them in a different way. There are the 3 degrees of freedom of the center of mass motion, and one degree of freedom for the kind of vibrational motion we saw in Figure (11-9) where we had two air carts connected by a spring. That accounts for 4 degrees of freedom; what are the other two?

When two connected particles move in three dimensions, what they can do that they could not do in one dimension is rotate about the center of mass. One can envision independent rotations about the x, the y, and the z axis, but one of these rotations does not count. In the picture we are developing, we will view the atoms themselves as perfectly smooth spheres, so that you cannot tell whether the atom itself is rotating or not. From this point of view, if the separation of two atoms in a diatomic molecule is along the z axis as shown in Figure (27), then rotation about the z axis cannot be detected and does not count as one of the degrees of freedom. Only rotations about the x and y axis contribute. Thus for a diatomic molecule moving in 3 dimensions, the 6 available degrees of freedom are 3 for center of mass motion, 2 for rotation, and one for vibration.



#### Figure 27

The three independent rotations of a molecule. For a diatomic molecule, the rotation about the z axis does not count. (Picture the atoms as perfectly smooth spheres. Then a collision could not start the z axis rotation, and you could not tell that it was rotating this way.)

If we have a system of n small spherical particles connected by spring-like forces, a reasonable model for many molecules, there should be 3n degrees of freedom. For example, the ammonia molecule with one nitrogen and three hydrogen atoms shown in Figure (28) should have 12 degrees of freedom. Three are center of mass motion, and now all 3 degrees of rotation should be counted. The remaining 6 must be vibrational normal modes, one for each spring like force. If you could kick an ammonia (or, let us say, a large scale model of one), record the motion of one of the molecules, and Fourier analyze a repeated pattern of the motion, you should be able to detect up to 6 normal mode frequencies of vibration.

We are now ready to state the equipartition of energy theorem that was first derived by James Clark Maxwell in 1858. Using Newtonian mechanics and the mathematical laws of probability, Maxwell showed that if a gas of molecules is in thermal equilibrium, then the average thermal energy of each molecule is 1/2 kT times the number of degrees of freedom possessed by the molecule. The theorem implies that, as you add thermal energy to a system of molecules, the energy is shared equally, on the average, between the available degrees of freedom.



#### Figure 28

The ammonia molecule is a tetrahedral structure with one nitrogen atom and three hydrogen atoms. Here we are modeling the forces between atoms as spring forces. The theorem is particularly easy to understand for the case of a gas of monatomic particles. If we have a single particle moving in 3 dimensions, we can write its kinetic energy  $1/2 \text{ mv}^2$  in the form  $1/2 \text{ m}(v_x^2 + v_y^2 + v_z^2)$  where we used the Pythagorean theorem to express  $v^2$  in terms of its components. Thus the kinetic energy breaks up into 3 distinct terms

kinetic  
energy = 
$$\frac{1}{2}mv_x^2 + \frac{1}{2}mv_y^2 + \frac{1}{2}mv_z^2$$

which we can call the kinetic energies of x motion, y motion, and z motion respectively.

If the particle is in thermal equilibrium, then on the average  $v_x^2$  should be the same as  $v_y^2$  and  $v_z^2$ . Thus the kinetic energies associated with each of the degrees of freedom of the molecule (x motion, y motion, and z motion) should be the same, and the sum should be the total average kinetic energy 3/2 kT. If 3/2 kT is shared 3 ways, each degree of freedom should get 1/2 kT kinetic energy on the average, as required by the equipartition of energy theorem.

## **Real Molecules**

Applying Maxwell's equipartition of energy theorem, we can make definite predictions about the specific heat of various kinds of molecules. We will begin with a brief review of the calculation of the specific heat of a monatomic gas like helium. A monatomic gas atom has 3 degrees of freedom, thus on the average its kinetic energy is

$$E(1 \text{ molecule}) = 3 \times \left(\frac{1}{2}kT\right) = \frac{3}{2}kT \qquad (34)$$

Since the specific heat  $C_V$  deals with one mole of a substance, we multiply Equation 34 through by Avagadro's number  $N_A$  (number of particles in a mole) to get

$$E(1 \text{ mole}) = N_{A}\left(\frac{3}{2}kT\right) = \frac{3}{2}(N_{A}k)T$$
$$= \frac{3}{2}RT \qquad (35)$$

where  $N_A k = R$  is the gas constant.

Finally  $C_V$  is defined as the change in energy  $\Delta E$  for a small change in temperature  $\Delta T$ . Differentiating Equation (35) gives  $\Delta E = (3/2)R\Delta T$ , so that

$$\frac{\Delta E}{\Delta T} \equiv C_V = \frac{3}{2}R \tag{36}$$

where the subscript V reminds us to keep the volume V of the gas constant so that none of the gas energy goes into doing the work of expanding the gas.

In the above derivation, we immediately see that the factor of 3 in the formula  $C_V = 3/2$  R came from our assumption that the molecule has 3 degrees of freedom. If we had a molecule with n degrees of freedom, then the equipartition of energy theorem predicts that the specific heat should be

for a molecule	prediction of the
with n degrees $C_V = \frac{n}{2}R$	equipartition of (37)
of freedom <sup>2</sup>	energy theorem

To see how good the predictions of the equipartition of energy theorem are, we have in Table 2 listed the specific heats of some common gasses, and compared the results with the predicted values.

# FAILURE OF CLASSICAL PHYSICS

If you worked out a complex theory that made detailed predictions, and when you compared the predictions with experiment, you got the results shown in Table 1, you should be disappointed. The agreement is simply terrible. The predictions work only for the monatomic gases (gases that remain individula atoms and do not form molecules). There is an increase in specific heat when we go to larger molecules, but not the predicted increase. In going from carbon dioxide to methane, where there is a considerable increase in the number of degrees of freedom, there is actually a decrease in the specific heat.

Maxwell worked on this problem for a number of years, carefully checking that he had correctly applied the mathematical laws of statistics to Newtonian mechanics, but he could find no error in his work. By 1879 he became convinced that Newtonian mechanics was flawed, and that eventually some new theory would have to be developed to replace it. The new theory, of course, was quantum mechanics, discovered nearly 50 years later. Maxwell's work in the 1860s and 1870s provided the first real evidence that Newtonian mechanics was not correct in all applications.

Molecule	Number of particles	Expected number degrees of freedom	Expected C <sub>V</sub> (joules/mole)	Experimental C <sub>V</sub>
helium	1	3	3/2 R = 12.5	12.5
argon	1	3	3/2 R = 12.5	12.6
nitrogen N <sub>2</sub>	2	6	6/2 R = 25	20.7
oxygen O <sub>2</sub>	2	6	6/2 R = 25	20.8
carbon dioxide CO <sub>2</sub>	3	9	9/2 R = 37.5	29.7
methane NH <sub>4</sub>	5	15	15/2R = 62.5	29.0

## Table 2

Specific heats of various molecules. Theory and experiment agree only for the monatomic gases.

## Freezing Out of Degrees of Freedom

While a straight forward application of the equipartition of energy theorem fails miserably, the ideas involved are not completely useless. A look at the specific heat of hydrogen gas as a function of temperature, Figure (29) gives us a clue as to what is happening at low temperatures below 100 K. The specific heat is 3/2 R, which is what we would expect for a monatomic gas with only 3 degrees of freedom. At these temperatures none of the thermal energy is going into exciting the internal motion of the atoms. At these low temperatures, the hydrogen molecules are acting like incompressible hard spheres.

Up at room temperature, the specific heat of hydrogen has jumped to 5/2 R. It appears that two additional degrees of freedom have appeared, and some of the thermal energy is now going into internal motions of the molecule. At still higher temperatures, just as the molecules are being torn apart, their specific heat reaches 7/2 R, indicating 7 degrees of freedom, one more than we expected. We can explain the 7 degrees of freedom by assuming that 1/2 kT of thermal energy goes, on the average, into the spring potential energy.

Going back down in temperature, we have the following picture. At very high temperature all the degrees of freedom are active and energy is shared equally among them as required by the equipartition of energy theorem. As we go down in temperature some of the degrees of freedom appear to freeze out. By room temperature we have lost two degrees of freedom, and down at 100 K, only the 3 translational degrees of freedom are left. Why the degrees of freedom freeze out is what is not explained by Newtonian mechanics. This is purely a quantum mechanical effect.

In Maxwell's time, the idea that matter consisted of atoms was a hypothesis rather than an experimentally proved fact. What atoms consisted of, whether they were indivisible hard spheres or had an internal structure was unknown. By applying Newtonian mechanics to models of atoms and molecules, he was trying to learn about the nature of these objects. The fact that monatomic gases have a specific heat  $C_V = 3/2R$ was evidence that the atoms were in fact acting like hard, indivisible objects. The failure to predict molecular specific heats turned out to be evidence that Newtonian mechanics was failing.

We know that atoms themselves consist of many particles—a nucleus surrounded by electrons. If we applied Newtonian mechanics to this structure, we would assume that each atom should have many degrees of freedom and that the nucleus and electrons should individually pick up thermal energy as the atom is heated. This simply does not happen. Applying the language we have used above, we can say that temperatures at which we ordinarily study atoms, the internal degrees of freedom of the atom are frozen out.

#### Figure 29

Specific heat of the hydrogen molecule. If each degree of freedom contributes 1/2R to the specific heat, then as the temperature drops, we see that various degrees of freedom freeze out. (Diagram adapted from Halliday and Resnick.)



## THERMAL EXPANSION

When you heat a substance, in most cases the substance expands. For example, the mercury (or alcohol) in the bulb at the bottom of a thermometer expands when heated, forcing more mercury up the small tube above the bulb. If the column is marked off in degrees centigrade, you have a typical mercury thermometer as illustrated in Figure (30). As another example, you are aware of the cracks left between sections of cement in sidewalks and cement highways. These cracks are there so that on a hot day when the cement expands, the sidewalk or road will not buckle.

The reason for thermal expansion can be understood at an atomic level in terms of the shape of the molecular force potential well. Figure (31) is a redrawing of the molecular force potential well of Figure (16), with some added information. Let 2r be the separation of the two atoms as indicated at the top of the diagram. If the molecule is at a very low temperature, the atom will essentially sit at the bottom of the potential well and the separation will be  $2r_0$  as shown.

If we raise the temperature of the molecule, the atoms gain thermal kinetic energy whose average value is 3/2 kT. As a result they will move back and forth at a higher level in the potential well, a height we have indicated as level 1 in the diagram. Due to the shape of



#### Figure 30

The typical mercury thermometer is based on the thermal expansion of mercury. There is no guarantee that a mercury thermometer and an ideal gas thermometer will agree at any temperatures except  $0^{\circ}C$  and  $100^{\circ}C$ . the potential energy well, due to the fact that the repulsive core rises faster than the attractive side, the average separation  $2r_T$  of the atoms at a temperature T is greater than the average separation  $2r_0$  at low temperatures.

Although our discussion of molecular forces focused on two atom molecules, the general shape of the molecular force potential well is the same when you have many atoms forming a liquid or a solid. Thus when you heat a liquid or a solid, the atoms gain an average thermal kinetic energy 3/2 kT, effectively rise up in the molecular force potential well, and due to the shape of the well, have a slightly greater average separation. The substance expands.

You can see that the amount of expansion depends upon the detailed shape of the potential well which varies from one substance to another. Thus a thermometer based on the expansion properties of mercury does not have to give precisely the same reading as a thermometer using alcohol, except at the calibration points  $0^{\circ}$  C and  $100^{\circ}$  C. And neither of these thermometers has to agree with the ideal gas thermometer. Since the ideal gas thermometer is based on the universal ideal gas law, one should use the ideal gas thermometer as a standard against which you calibrate mercury, alcohol, and other thermometers.



#### Figure 31

As you raise the temperature of a substance, the average thermal kinetic energy 3/2 kT rises and the molecule sits higher in its potential well. Because the well is lopsided, the average separation of the atom becomes greater and the substance expands.

# **OSMOTIC PRESSURE**

We would like to conclude this chapter, this brief view of atoms, molecules and thermal processes, with a discussion of two familiar phenomena that can be understood qualitatively from a molecular point of view. One is the elasticity of rubber, and the other is the process of osmosis, which is essential for biological systems.

Osmosis is a rather peculiar but important effect that is easily explained with an atomic model. Ordinarily, when a liquid can flow between two vessels at the same height, the liquid will tend to seek the same height in both vessels. But this does not always happen. Suppose we have a tank separated by a membrane, as shown in Figure (32). On the right side (side 2) of the membrane we place pure water, indicated by the small molecules. On the left (side 1) we place a solution of water and some other substance consisting of large molecules. The membrane has a special characteristic: the small water molecules can pass through it easily, whereas the big molecules are prevented from passing through because the holes in the membrane are too small. Initially, the two compartments are filled to the same level on each side of the membrane.

Assume that some definite fraction, for instance 50%, of all water molecules that strike the membrane pass through it. Initially, more water molecules strike the membrane from side 2 than from side 1 simply because there are more water molecules on side 2. If more water molecules strike from side 2 than side 1, and if 50% of **all** the water molecules striking the membrane pass through it, there must be a net flow of water from side 2 to side 1.

As the flow continues, the level of the liquid on side 1 rises and the solution becomes diluted. As the solution becomes further diluted, the number of water molecules on side 1 facing each cm<sup>2</sup> of the membrane increases, the more flow back to side 2, so that the net flow into side 1 decreases. From this description alone, however, we would not expect the flow to stop, since we never get pure water on side 1.

The flow does stop eventually though, because the level in side 1 rises to such an extent that the pressure at the bottom of side 1 becomes considerably greater than the pressure at the bottom of side 2 (a result of the increased weight of the column of water). This additional pressure, known as *osmotic pressure*, finally stops the flow of water from side 2 to side 1. The flow of the small molecules through the membrane is called *osmosis*; thus osmotic pressure is the pressure that finally stops osmosis.



## Figure 32

Osmosis. The two sides of the container are separated by a membrane that allows the water (small molecules) but not the large molecules to pass through. If the liquid levels are the same initially, as in (a), some of the pure water will flow through the membrane, raising the level on the side with the large molecules as shown in (b). This process is called osmosis.

Osmosis and osmotic pressure are crucial in biological processes. Osmosis is involved in the separation of nutrients and wastes in our own cells, the flow of fluids in our bodies, the flow of sap in plant life, and a number of other important processes.

The function and composition of blood is critically dependent on osmosis and osmotic pressure. Blood consists of red cells, white cells, and a fluid called *plasma*. The red cells are membrane sacs containing about 60% water and 40% hemoglobin molecules, molecules closely related to, but about four times as large as, the giant myoglobin molecule described in at the beginning of this chapter. We may think of the red blood cell as representing side 1 in Figure (32) where the big molecules are the hemoglobin molecules. If red blood cells are removed from blood and placed in pure water, they absorb so much water by osmosis that they burst. The red hemoglobin flows away, leaving an empty, pale misshapen sac.

The function of the red blood cell and its hemoglobin is to carry oxygen to the other cells in the body. The plasma, which consists of 90% water, 9% protein molecules, and 1% salts, serves as a fluid in which to dissolve needed proteins and salts to be carried to the cells, and to make the blood fluid enough to flow through the minute capillaries.

The capillary walls through which blood flows are porous membranes that permit water and salts to pass freely through, but that restrict the passage of proteins. Pure water could not be pumped through the bloodstream because it would leak out through the capillary walls. You may wonder how blood plasma, which is 90% water, can be pumped through the porous capillaries. The reason is that the 9% protein molecules in the plasma is sufficient to draw just enough water back into the capillary by osmosis to replace the water molecules that do leak out. Just as many water molecules are drawn back in as leak out, even though the pressure of the plasma inside the capillary is greater than the pressure of the fluids outside the leaky walls. Thus, side 1 in Figure (32) behaves in the same way as the capillary with the blood plasma inside it.

# **ELASTICITY OF RUBBER**

A model for the elasticity of rubber was presented by Richard Feynman in a lecture to freshmen at Caltech in 1960. We select this model, not so much for its accuracy in describing the detailed behavior of molecules in rubber, but for developing an intuition for thermal processes. The mechanisms underlying the model and the behavior of rubber are fundamentally the same. The beauty of the model is that it is so outrageous that you are forced to think differently about thermal processes.

# A Model of Rubber

Imagine that you enter a large room where there are a number of heavy chains loosely suspended from one end of the room to the other, as shown in Figure (33). These are massive chains, like the anchor chains used on old sailing ships, but they are hanging loosely, so that except for their weight, they are not exerting any force pulling the walls together.

On the floor are hundreds of cannonballs, lying there a couple of layers deep. This is our room at "absolute zero".

Now turn up the temperature in the room. The cannonballs start to jiggle and vibrate with an average thermal kinetic energy 3/2 kT. In this model, nothing melts. Instead, as we turn up the temperature the jiggling becomes stronger and stronger. When the average thermal kinetic energy 3/2 kT becomes as large as the gravitational potential energy mgh of a cannonball near the ceiling, then we will have cannonballs flying all around the room. We will have a gas of cannonballs.

As the cannonballs fly around, they strike the chains, kinking them up as indicated in Figure (34). The kinked-up chains are no longer hanging loose, instead they are taut and pulling the side walls of the room in.

If we raise the temperature of the gas of cannonballs, the cannonballs strike the chains harder and the chains pull harder on the walls.

Here is an experiment that stretches the imagination even more. Suppose we start with the room with a gas of cannonballs at a temperature T, and chains kinked by



**Figure 33** *Room with suspended chains and cannonballs.* 

the colliding cannonballs, and suddenly pull the sides of the room apart so that the chains are straight and tight. When we suddenly straighten out the kinked chains, the chains will slap against the cannonballs transforming the work we do pulling the chains straight into increased thermal kinetic energy of the cannonballs. As a result by suddenly stretching the chains we raise the temperature of the cannonballs.

If we let the walls go back suddenly, the chains initially go slack, and it takes some of the thermal kinetic energy of the cannonballs to kink the chains up again. As a result of unstretching the chains, the temperature of the cannonballs drops.

One's lips are a good detector of small temperature changes. Place a loose rubber band between your lips and suddenly stretch it. You will notice that the rubber band becomes distinctly warmer. Now quickly release the rubber band by bringing your hands together. The rubber band becomes distinctly cool. Rubber consists of a long chain of molecules that are kinked up by thermal motion. When you stretch the rubber band, you increase the thermal kinetic energy of the molecules and raise their temperature. Releasing the band reduces the thermal motion and drops the temperature. The elastic restoring force you felt when you stretched the band is caused by thermal motions kinking the long chain molecules.



**Figure 34** *Top view of chains being struck by cannonballs.* 

# Chapter 18 Entropy



**Movie Undive** 

The focus of the last chapter was the thermal energy of the atoms and molecules around us. While the thermal energy of an individual molecule is not large, the thermal energy in a reasonable collection of molecules, like a mole, is a noticeable amount. Suppose, for example, you could extract all the thermal energy in a mole of helium atoms at room temperature. How much would that energy be worth at a rate of 10 cents per kilowatt hour?

This is an easy calculation to do. The atoms in helium gas at room temperature have an average kinetic energy of 3/2 kT per molecule, thus the energy of a mole is  $N_A(3/2KT) = 3/2RT$ , where  $N_A$  is Avagadro's number and  $R = N_A K$ . Since ice melts at 273 K and water boils at 373 K, a reasonable value for room temperature is 300 K, about one quarter of the way up from freezing to boiling. Thus the total thermal energy in a mole of room temperature helium gas is

thermal energy  
in a mole of  
room temperature  
helium gas 
$$= \frac{3}{2}RT$$
$$= \frac{3}{2} \times 8 \frac{joules}{mole K} \times 300K$$
$$= 3600 \text{ joules}$$

This is enough energy to lift 1 kilogram to a height of 360 feet!

To calculate the monetary value of this energy, we note that a kilowatt hour is 1000 watts of electric power for 3600 seconds, or  $1000 \times 3600$  joules. Thus at a rate of 10 cents per kilowatt hour, the thermal energy of a mole of helium gas is worth only .01 cents.

The value .01 cents does not sound like much, but that was the value of the energy in only one mole of helium. Most substances have a greater molar heat capacity than helium due to the fact that energy is stored in internal motions of the molecule. Water at room temperature, for example, has a molar heat capacity six times greater than that of helium. Thus we would expect that the thermal energy in a mole of water should be of the order of 6 times greater than that of helium, or worth about .06 cents.

A mole of water is only 18 grams. A kilogram of water, 1000 grams of it, is 55 moles, thus the thermal energy in a kilogram or liter of water should have a value in the neighborhood of .06 cents/mole  $\times$  55 moles = 3.3 cents. Now think about the amount of water in a swimming pool that is 25 meters long, 10 meters wide, and 2 meters deep. This is 500 m<sup>3</sup> or 500  $\times$ 10<sup>3</sup> liters, their being 1000 liters/m<sup>3</sup>. Thus the commercial value of the heat energy in a swimming pool of water at room temperature is 5  $\times$  10<sup>5</sup>  $\times$  3.3 cents or over \$16,000.

The point of this discussion is that there is a lot of thermal energy in the matter around us, energy that would have enormous value if we could get at it. The question is why don't we use this thermal energy rather than getting energy by burning oil and polluting the atmosphere in the process?

# INTRODUCTION

A simple lecture demonstration helps provide insight into why we cannot easily get at, and use, the \$16,000 of thermal energy in the swimming pool. In this demonstration, illustrated in Figure (1), water is pumped up through a hose and squirted down onto a flat plate placed in a small bucket as shown. If you use a vibrator pump, then the water comes out as a series of droplets rather than a continuous stream.

To make the individual water droplets visible, and to slow down the apparent motion, the water is illuminated by a strobe. If the time between strobe flashes is just a bit longer than the time interval which the drops are ejected, the drops will appear to move very slowly. This allows you to follow what appears to be an individual drop as it moves down toward the flat plate.

For our discussion, we want to focus on what happens to the drop as it strikes the plate. As seen in the series of pictures in Figure (2), when the drop hits it flattens out, creating a wave that spreads out from where the drop hits. The wave then moves down the plate into the pool of water in the bucket. [In Figure (1) we see several waves flowing down the plate. Each was produced by a separate drop.] Let us look at this process from the point of view of the energy involved. Before the drop hits, it has kinetic energy due to falling. When it hits, this kinetic energy goes into the kinetic energy of wave motion. The waves then flow into the bucket, eventually dissipate, and all the kinetic energy becomes thermal energy of the water molecules in the bucket. This causes a slight, almost undetectable, increase in the temperature of the water in the bucket (if the water drop had the same temperature as the water in the bucket, and we neglect cooling from evaporation).

We have selected this demonstration for discussion, because with a slight twist of the knob on the strobe, we can make the process appear to run backwards. If the time interval between flashes is just a bit shorter than the pulse interval of the pump, the drops appear to rise from the plate and go back into the hose. The situation looks funny, but it makes a good ending to the demonstration. Everyone knows that what they see couldn't possibly happen—or could it?

Does this reverse flow violate any laws of physics? Once a drop has left the plate it moves like a ball thrown up in the air. From the point of view of the laws of physics, nothing is peculiar about the motion of the drop from the time it leaves the plate until it enters the hose.

#### Figure 1

Water droplets are created by a vibrator pump. If you illuminate the drops with a strobe light, you can make them appear to fall or rise.





Where the situation looks funny is in the launching of the drop from the plate. But it turns out that none of the laws of physics we have discussed so far is violated there either. Let us look at this launching from the point of view of the energy involved. Initially the water in the bucket is a bit warm. This excess thermal energy becomes organized into a wave that flows up the plate. As seen in Figure (3) the wave coalesces into a drop that is launched up into the air. No violation of the law of conservation of energy is needed to describe this process.

While the launching of the drop in this reversed picture may not violate the laws of physics we have studied, it still looks funny, and we do not see such things happen in the real world. There has to be some reason why we don't. The answer lies in the fact that in the reversed process we have converted thermal energy, the disorganized kinetic energy of individual molecules, into the organized energy of the waves, and finally into the more concentrated kinetic energy of the upward travelling drop. We have converted a disorganized form of energy into an organized form in a way that nature does not seem to allow. It is not impossible to convert thermal energy into organized kinetic energy or what we call useful work. Steam engines do it all the time. In a modern electric power plant, steam is heated to a high temperature by burning some kind of fuel, and the steam is sent through turbines to produce electricity. A certain fraction of the thermal energy obtained from burning the fuel ends up as electrical energy produced by the electric generators attached to the turbines. This electric energy can then be used to do useful work running motors.

The important point is that power stations cannot simply suck thermal energy out of a reservoir like the ocean and turn it into electrical energy. That would correspond to our water drop being launched by the thermal energy of the water in the bucket.



**Figure 2** *Falling water drop creating wave.* 



Figure 3 Rising wave launching water drop.

Even more discouraging is the fact that power plants do not even use all the energy they get from burning fuel. A typical high efficiency power plant ends up discarding, into the atmosphere or the ocean, over 2/3 of the energy it gets from burning fuel. Less than 1/3 of the energy from the fuel is converted into useful electrical energy. Car engines are even worse. Less than 1/5 of the energy from the gasoline burned goes into powering the car; most of the rest comes out the exhaust pipe.

Why do we tolerate these low efficiency power plants and even lower efficiency car engines? The answer lies in the problem of *converting a disorganized form of energy into an organized one*. Or to state the problem more generally, of trying to create order from chaos.

The basic idea is that a disorganized situation does not naturally organize itself—in nature, things go the other way. For example, if you have a box of gas, and initially the atoms are all nicely localized on one side of the box, a short time later they will be flying around throughout the whole volume of the box. On their own, there is almost no chance that they will all move over to that one side again. If you want them over on one side, you have to do some work, like pushing on a piston, to get them over there. It takes work to create order from disorder.

At first, it seems that the concepts of order and disorder, and the related problems of converting thermal energy into useful work should be a difficult subject to deal with. If you wished to formulate a physical law, how do you go about even defining the concepts. What, for example, should you use as an experimental definition of disorder? It turns out, surprisingly, that there is a precise definition of a quantity called *entropy* which represents the amount of disorder contained in a system. Even more surprising, the concept of entropy was discovered before the true nature of heat was understood.

The basic ideas related to entropy were discovered in 1824 by the engineer Sadi Carnot who was trying to figure out how to improve the efficiency of steam engines. Carnot was aware that heat was wasted in the operation of a steam engine, and was studying the problem in an attempt to reduce the waste of heat. In his studies Carnot found that there was a *theoretically* maximum efficient engine whose efficiency depended upon the temperature of the boiler relative to the temperature of the boiler's surroundings.

To make his analysis, Carnot had to introduce a new assumption not contained in Newton's law of mechanics. Carnot's assumption is equivalent to the idea that *you cannot convert thermal energy into useful work in a process involving only one temperature*. This is why you cannot sell the \$16,000 worth of thermal energy in the swimming pool—you cannot get it out.

This law is known as the *Second Law of Thermodynamics*. (The first law is the law of conservation of energy itself.) The second law can also be expressed in terms of entropy which we now know represents the disorder of a system. The second law states that in any process, the total entropy (disorder) of a system either stays the same or increases. Put another way, it states that in any process, the total order of a system cannot increase; it can only stay the same, or the system can become more disordered.

To develop his formulas for the maximum efficiency of engines, Carnot invented the concept known as a *Carnot engine*, based on what is called the *Carnot cycle*. The Carnot engine is not a real engine, no one has ever built one. Instead, you should think of it as a *thought experiment*, like the ones we used in Chapter 1 to figure out what happened to moving clocks if the principle of relativity is correct.

The question we wish to answer is, how efficient can you make an engine or a power plant, if the second law of thermodynamics is correct? If you cannot get useful work from thermal energy at one temperature, how much work can you get if you have more than one temperature? It turns out that there is a surprisingly simple answer, but we are going to have to do quite a bit of analysis of Carnot's thought experiment before we get the answer. During the discussion of the Carnot engine, one should keep in mind that we are making this effort to answer one basic question—what are the consequences of the second law of thermodynamics what are the consequences of the idea that order does not naturally arise from disorder.

# WORK DONE BY AN EXPANDING GAS

The Carnot thought experiment is based on an analysis of several processes involving the ideal piston and cylinder we discussed in the last chapter. We will discuss each of these processes separately, and then put them together to complete the thought experiment.

The ideal piston and cylinder is shown in Figure (4). A gas, at a pressure p, is contained in the cylinder by a frictionless piston of cross-sectional area A. (Since no one has yet built a piston that can seal the gas inside the cylinder and still move frictionlessly, we are now already into the realm of a thought experiment.) A force F is applied to the outside of the piston as shown to keep the piston from moving. The gas, at a pressure p, exerts an outward force

$$p\left(\frac{newtons}{meter^2}\right) \times A(meter^2) = pA(newtons)$$

on the cylinder, thus F must be given by

F=pA (1)

to keep the cylinder from moving.

If we decrease the force F just a bit to allow the gas in the cylinder to expand, the expanding gas will do work on the piston. This is because the gas is exerting a force pA on the cylinder, while the cylinder is moving in the direction of the force exerted by the gas. If the piston moves out a distance  $\Delta x$  as shown in Figure (5), the work  $\Delta W$  done by the gas is the force pA it exerts times the distance  $\Delta x$ 



#### Figure 4

In the ideal piston and cylinder, the piston confines the gas and moves frictionlessly.

$$\Delta W = p\Delta V \tag{3}$$

Equation 3 is more general than our derivation indicates. Any time a gas expands its volume by an amount  $\Delta V$ , the work done by the gas is  $(p\Delta V)$  no matter what the shape of the container. For example, if you heat the gas in a balloon and the balloon expands a bit, the work done by the gas is  $(p\Delta V)$  where  $\Delta V$  is the increase in the volume of the balloon.

#### Exercise 1

In our introduction to the concept of pressure, we dipped a balloon in liquid nitrogen until the air inside became a puddle of liquid air (see Figure 17-19). When we took the balloon out of the liquid nitrogen, the air slowly expanded until the balloon returned to its original size. During the expansion, the rubber of the balloon was relatively loose, which means that the air inside the balloon remained at or very near to atmospheric pressure during the entire time the balloon was expanding.

(a) If the final radius of the balloon is 30 cm, how much work did the gas inside the balloon do as the balloon expanded? (You may neglect the volume of the liquid air present when the expansion started.) (Answer:  $1.1 \times 10^4$  joules.)

(b) Where did the gas inside the balloon get the energy required to do this work?



The work done by an expanding gas is equal to the force pA it exerts, times the distance  $\Delta x$  the piston moves.
# SPECIFIC HEATS C<sub>v</sub> AND C<sub>p</sub>

In our earlier discussion of specific heat, we dealt exclusively with the "molar" specific heat at constant volume  $C_V$ . We always assumed that we kept the gas at constant volume so that all the energy we added would go into the internal energy of the gas. If we had allowed the gas to expand, then some of the energy would have gone into the work the gas did to expand its volume, and we would not have had an accurate measure of the amount of energy that went into the gas itself.

Sometimes it is convenient to heat a gas while keeping the gas *pressure*, rather than *volume*, constant. This is more or less the case when we heat the gas in a balloon. The balloon expands, but the pressure does not change very much if the expansion is small.

Earlier we defined the molar heat capacity  $C_V$  as the amount of energy required to heat one mole of a substance one kelvin, if the volume of the substance is kept constant. Let us now define the molar heat capacity  $C_p$  as the amount of energy required to heat one mole of a substance one kelvin if the pressure is kept constant. For gases  $C_p$  is always larger than  $C_V$ . This is because, when we heat the gas at constant pressure, the energy goes both into heating and expanding the gas. When we heat the gas at constant volume, the energy goes only into heating the gas. We can write this out as an equation as follows



Noting that since  $C_V$  is equal to the increase in thermal energy of the gas, and that the work done is  $p\Delta V$ , we get

$$C_{p} = C_{V} + p\Delta V \tag{5}$$

In the special case of an ideal gas, we can use the ideal gas law pV = nRT, setting n = 1 for 1 mole

If we let the gas expand a bit at constant pressure, we get differentiating Equation 6, keeping p constant\*\*

$$p\Delta V = R\Delta T$$
 (if p is constant) (7)

\*\* By differentiating the equation (pV = RT), we mean that we wish to equate the *change* in (pV) to the *change* in (RT). To determine the change in (pV), for example, we let (p) go to  $(p + \Delta p)$  and (V) go to  $(V + \Delta V)$ , so that the product (pV) becomes

$$pV \rightarrow (p + \Delta p)(V + \Delta V)$$
  
= pV + (\Delta p)V + p\Delta V + (\Delta p)\Delta V

If we neglect the second order term  $(\Delta p)\Delta V$  then

$$pV \rightarrow pV + (\Delta p)V + p\Delta V$$

Then if we hold the pressure constant ( $\Delta p = 0$ ), we see that the change in (pV) is simply (p $\Delta V$ ). Since R is constant, the change in RT is simply R $\Delta T$ .

If the temperature increase is  $\Delta T = 1$  kelvin, then Equation 7 becomes

$$p\Delta V = R$$
 (1 mole, p constant,  $\Delta T = 1K$ ) (8)

Using Equation 8 in Equation 5 we get the simple result

$$C_{p} = C_{V} + R \tag{9}$$

The derivation of Equation 9 illustrates the kind of steps we have to carry out to calculate what happens to the heat we add to substances. For example, in going from Equation 6 to Equation 7, we looked at the change in volume when the temperature but not the pressure was varied. When we make infinitesimal changes of some quantities in an equation while holding the quantities constant, the process is called *partial differentia-tion*. In this text we will not go into a formal discussion of the ideas of partial differentiation. When we encounter the process, the steps should be fairly obvious as they were in Equation 7.

(The general subject that deals with changes produced by adding or removing heat from substances is called **thermodynamics**. The full theory of thermodynamics relies heavily on the mathematics of partial derivatives. For our discussion of Carnot's thought experiments, we need only a small part of thermodynamics theory.)

#### **Exercise 2**

(a) Back in Table 2 on page 31 of Chapter 17, we listed the values of the molar specific heats for a number of gases. While the experimental values of did not agree in most cases with the values predicted by the equipartition of energy, you can use the experimental values of  $C_v$  to accurately predict the values of  $C_p$  for these gases. Do that now.

(b) Later in this chapter, in our discussion of what is called the *adiabatic* expansion of a gas (an expansion that allows no heat to flow in), we will see that the ratio of  $C_p/C_V$  plays an important role in the theory. It is common practice to designate this ratio by the Greek letter  $\gamma$ 

$$\gamma \equiv C_{\rm p} / C_{\rm V} \tag{10}$$

(i) Explain why, for an ideal gas,  $\gamma$  is always greater than 1.

(ii) Calculate the value of  $\gamma$  for the gases, listed in Table 2 of Chapter 17.

Answers:

gas	γ
helium	1.66
argon	1.66
nitrogen	1.40
oxygen	1.40
CO <sub>2</sub>	1.28
$NH_4$	1.29

# ISOTHERMAL EXPANSION AND PV DIAGRAMS

In the introduction, we pointed out that while a swimming pool of water may contain \$16,000 of thermal energy, we could not extract this energy to do useful work. To get useful energy, we have to burn fuel to get heat, and convert the heat to useful work. What seemed like an insult is that even the most efficient power plants turn only about 1/3 of the heat from the fuel into useful work, the rest being thrown away, expelled either into the atmosphere or the ocean.

We are now going to discuss a process in which heat is converted to useful work with 100% efficiency. This involves letting the gas in a piston expand at constant temperature, in a process called an *isothermal expansion*. (The prefix "iso" is from the Greek meaning "equal," thus, isothermal means equal or constant temperature.) This process cannot be used by power plants to make them 100% efficient, because the process is not repetitive. Some work is required to get the piston back so that the expansion can be done over again.

Suppose we start with a gas in a cylinder of volume  $V_1$  and let the gas slowly expand to a volume  $V_2$  as shown in Figure (6). We control the expansion by adjusting the force  $\vec{F}$  exerted on the back side of the piston.

While the gas is expanding, it is doing work on the piston. For each  $\Delta V$  by which the volume of the gas increases, the amount of work done by the gas is  $p\Delta V$ . The energy required to do this work must come from somewhere. If we did not let any heat into the cylinder, the energy would have to come from thermal energy, and the temperature would drop. (This is one way to get work out of thermal energy.)



**Figure 6** Isothermal expansion of the gas in a cylinder. The force  $\vec{F}$  on the cylinder is continually adjusted so that the gas expands slowly at constant temperature.

However, we wish to study the process in which the gas expands at constant temperature. To keep the temperature from dropping, we have to let heat flow into the gas. Since the temperature of the gas is constant, there is no change in the thermal energy of the gas. Thus *all the heat that flows in goes directly into the work done by the gas*.

To calculate the amount of work done, we have to add up all the  $p\Delta V$ 's as the gas goes from a volume  $V_1$  to a volume  $V_2$ . If we graph pressure as a function of volume, in what is called a *pV diagram*, we can easily visualize these increments of work  $p\Delta V$  as shown in Figure (7).

Suppose the pressure of the gas is initially  $p_1$  when the volume of the cylinder is  $V_1$ . As the cylinder moves out and the gas expands, its pressure will drop as shown in Figure (7), reaching the lower value  $p_2$  when the cylinder volume reaches  $V_2$ . At each step  $\Delta V_i$ , when the pressure is  $p_i$ , the amount of work done by the gas is  $p_i \Delta V_i$ . The total work, the sum of all the  $p_i \Delta V_i$ , is just the total area under the pressure curve, as seen in Figure (7).

The nice feature of a graph of pressure versus volume like that shown in Figure (7), is *the work done by the gas is always the area under the pressure curve*, no matter what the conditions of the expansion are. If we had allowed the temperature to change, the shape of the pressure curve would have been different, but the work done by the gas would still be the area under the pressure curve.



Figure 7

The work done by an expanding gas is equal to the sum of all  $p \Delta V$ 's, which is the area under the pressure curve.

#### Isothermal Compression

If we shoved the piston back in, from a volume  $V_2$  to a volume  $V_1$  in Figure (7), we would have to do work on the gas. If we kept the temperature constant, then the pressure would increase along the curve shown in Figure (7) and the work we did would be precisely equal to the area under the curve. In this case work is done on the gas (we could say that during the compression the gas does negative work). When work is done on the gas, the temperature of the gas will rise unless we let heat flow out of the cylinder. Thus if we have an isothermal compression, where there is no increase in the thermal energy of the gas, then we have the pure conversion of useful work into the heat expelled by the piston. This is the opposite of what we want for a power plant.

#### **Isothermal Expansion of an Ideal Gas**

If we have one mole of an ideal gas in our cylinder, and keep the temperature constant at a temperature  $T_1$ , then the gas will obey the ideal gas equation.

$$pV = RT_1 = constant$$
 (11)

Thus the equation for the pressure of an ideal gas during an isothermal expansion is

$$p = \frac{constant}{V}$$
(1la)

and we see that the pressure decreases as 1/V. This decrease is shown in the pV diagram of Figure (8).



In the isothermal expansion of an ideal gas, we have pV = constant. Thus the pressure decreases as 1/V.

#### ADIABATIC EXPANSION

We have seen that we can get useful work from heat during an isothermal expansion of a gas in a cylinder. As the gas expands, it does work, getting energy for the work from heat that flows into the cylinder. This represents the conversion of heat energy at one temperature into useful work. The problem is that there is a limited amount of work we can get this way. If we shove the piston back in so that we can repeat the process and get more work, it takes just as much work to shove the piston back as the amount of work we got out during the expansion. The end result is that we have gotten nowhere. We need something besides isothermal expansions and compressions if we are to end up with a net conversion of heat into work.

Another kind of expansion is to let the cylinder expand without letting any heat in. This is called an *adiabatic expansion*, where *adiabatic* is from the Greek (*a*-not + *dia-through* + *bainein*-to go). If the gas does work during the expansion, and we let no heat energy in, then *all the work must come from the thermal energy* of the gas. The result is that the gas will cool during the expansion. In an adiabatic expansion, we are converting the heat energy *contained in the gas* into useful work. If we could keep this expansion going we could suck all the thermal energy out of the gas and turn it into useful work. The problem, of course, getting the piston back to start the process over again. It is instructive to compare an isothermal expansion to an adiabatic expansion of a gas. In either case the pressure drops. But in the adiabatic expansion, the pressure drops faster because the gas cools. In Figure (9), we compare the isothermal and adiabatic expansion curves for an ideal gas. Because the adiabatic curve drops faster in the pV diagram, there is less area under the adiabatic curve, and the gas does less work. This is not too surprising, because less energy was available for the adiabatic expansion since no heat flowed in.

For an ideal gas, the equation for an adiabatic expansion is

$$pV^{\gamma} = constant; \quad \gamma = \frac{C_p}{C_V}$$
 (12)

a result we derive in the appendix. (You calculated the value of  $\gamma$  for various gases in Exercise 2.) The important point now is not so much this formula, as the fact that the adiabatic curve drops faster than the isothermal curve.

If we compress a gas adiabatically, all the work we do goes into the thermal energy of the gas, and the temperature rises. Thus with an adiabatic expansion we can lower the temperature of the gas, and with an adiabatic compression raise it.



#### Figure 9

Comparison of isothermal and adiabatic expansions. In an adiabatic expansion the gas cools, and thus the pressure drops faster.

#### **Exercise 3**

In the next section, we will discuss a way of connecting adiabatic and isothermal expansions and compressions in such a way that we form a complete cycle (get back to the starting point), and get a net amount of work out of the process. Before reading the next section, it is a good exercise to see if you can do this on your own.

In order to see whether or not you are getting work out or putting it in, it is useful to graph the process in a pV diagram, where the work is simply the area under the curve.

To get you started in this exercise, suppose you begin with an ideal gas at a pressure  $p_1$ , volume  $V_1$ , and temperature  $T_1$ , and expand it isothermally to  $p_2$ ,  $V_2$ ,  $T_1$  as shown in Figure (10a). The work you get out is the area under the curve.

If you then compressed the gas isothermally back to  $p_1$ ,  $V_1$ ,  $T_1$ , this would complete the cycle (get you back to where you started), but it would take just as much work to compress the gas as you got from the expansion. Thus there is no net work gained from this cycle. A more complex cycle is needed to get work out.

If we add an adiabatic expansion to the isothermal expansion as shown in Figure (10b) we have the start of something more complex. See if you can complete this cycle, i.e., get back to  $p_1$ ,  $V_1$ ,  $T_1$ , using adiabatic and isothermal expansions or compressions, and get some net work in the process. See if you can get the answer before we give it to you in the next section. Also show graphically, on Figure (10b) how much work you do get out.



**Figure 10a** *pV diagram for an isothermal expansion from volume*  $V_1$  *to volume*  $V_2$ .

# THE CARNOT CYCLE

With the isothermal and adiabatic expansion and compression of an ideal gas in a frictionless cylinder, we now have the pieces necessary to construct a Carnot cycle, the key part of our thought experiment to study the second law of thermodynamics.

The goal is to construct a device that continually converts heat energy into work. Such a device is called an *engine*. Both the isothermal and adiabatic expansions of the gas converted heat energy into work, but the expansions alone could not be used as an engine because the piston was left expanded. Carnot's requirement for an engine was that after a complete cycle all the working parts had to be back in their original condition ready for another cycle. Somehow the gas in the cylinder has to be compressed again to get the piston back to its original position. And the compression cannot use up all the work we got from the expansion, in order that we get some net useful work from the cycle.

The idea for Carnot's cycle that does give a net amount of useful work is the following. Start off with the gas in the cylinder at a high temperature and let the gas expand isothermally. We will get a certain amount of work from the gas. Then rather than trying to compress the hot gas, which would use up all the work we got, cool the gas to reduce its pressure. Then isothermally compress the cool gas. It should take less work to compress the low pressure cool gas than the work we got from the high pressure hot gas. Then finish the



**Figure 10b** *pV diagram for an isothermal expansion followed by an adiabatic expansion.* 

cycle by heating the cool gas back up to its original temperature. In this way you get back to the original volume and temperature (and therefore pressure) of the cylinder; you have a complete cycle, and hopefully you have gotten some useful work from the cycle.

To cool the gas, and then later heat it up again, Carnot used an adiabatic expansion and then an adiabatic compression. We can follow the steps of the Carnot cycle on the pV diagram shown in Figure (11). The gas starts out at the upper left hand corner at a high temperature  $T_1$ , volume  $V_1$ , pressure  $p_1$ . It then goes through an isothermal expansion from a volume  $V_1$  to a volume  $V_2$ , remaining at the initial temperature  $T_1$ . The hot gas is then cooled down to a low temperature  $T_3$  by an adiabatic expansion to a volume  $V_3$ . The cool, low pressure gas is then compressed isothermally to a volume  $V_4$ , where it is then heated back to a higher temperature  $T_1$  by an adiabatic compression. The volume  $V_4$  is chosen just so that the adiabatic compression will bring the temperature back to  $T_1$  when the volume gets back to  $V_1$ .







The Carnot Cycle. The gas first expands at a high temperature  $T_1$ . It is then cooled to a lower temperature  $T_3$  by an adiabatic expansion. Then it is compressed at this lower temperature, and finally heated back to the original temperature  $T_1$  by an adiabatic compression. We get a net amount of work from the process because it takes less work to compress the cool low pressure gas than we got from the expansion of the hot high pressure gas.

In this set of 4 processes, we get work out of the two expansions, but put work back in during the two compressions. Did we really get some net work out? We can get the answer immediately from the pV diagram. In Figure (12a), we see the amount of work we got out of the two expansions. It is the total area under the expansion curves. In Figure (12b) we see how much work went back in during the two compressions. It is the total area under the two compressions. It is the total area under the two compression curves. Since there is more area under the expansion curves than the compression curves, we got a net amount of work out. The net work out is, in fact, just equal to the 4 sided area between the curves, seen in Figure (13).



Figure 12a

The work we get out of the two expansions is equal to the area under the expansion curves.





The work required to compress the gas back to its original volume is equal to the area under the compression curves.

# Thermal Efficiency of the Carnot Cycle

The net effect of the Carnot cycle is the following. During the isothermal expansion while the cylinder is at the high temperature  $T_H$ , a certain amount of thermal or heat energy, call it  $Q_H$ , flows into the cylinder.  $Q_H$  must be equal to the work the gas is doing during the isothermal expansion *since the gas' own thermal energy does not change at the constant temperature*. (Here, all the heat in becomes useful work.)

During the isothermal compression, while the cylinder is at the lower temperature  $T_L$ , (the gas having been cooled by the adiabatic expansion), an amount of heat  $Q_L$  is expelled from the cylinder. Heat must be expelled because we are doing work on the gas by compressing it, and *none of the energy we supply can go into the thermal energy of the gas* because its temperature is constant. (Here all the work done becomes expelled heat.)

Since no heat enters or leaves the cylinder during the adiabatic expansion or compression, all flows of heat have to take place during the isothermal processes. Thus the net effect of the process is that an amount of thermal energy or heat  $Q_H$  flows into the cylinder at the high temperature  $T_H$ , and an amount of heat  $Q_L$  flows out at the low temperature  $T_L$ , and we get a net amount of useful work W out equal to the 4-sided area seen in Figure (13). By the law of conservation of energy, the



#### Figure 13

The net work we get out of one complete cycle is equal to the area bounded by the four sided shape that lies between the expansion and compression curves. work W must be equal to the difference between  $Q_H$  in and  $Q_L$  out

$$W = Q_H - Q_L \tag{13}$$

We see that the Carnot engine suffers from the same problem experienced by power plants and automobile engines. They take in heat  $Q_H$  at a high temperature (produced by burning fuel) and do some useful work W, but they expel heat  $Q_L$  out into the environment. To be 100% efficient, the engine should use all of  $Q_H$  to produce work, and not expel any heat  $Q_L$ . But the Carnot cycle does not appear to work that way.

One of the advantages of the Carnot cycle is that we can calculate  $Q_H$  and  $Q_L$ , and see just how efficient the cycle is. It takes a couple of pages of calculations, which we do in the appendix, but we obtain a remarkably simple result. The ratio of the heat in,  $Q_H$ , to the heat out,  $Q_L$ , is simply equal to the ratio of the high temperature  $T_H$  to the low temperature  $T_L$ .

$$\left| \begin{array}{c} \frac{Q_{H}}{Q_{L}} = \frac{T_{H}}{T_{L}} \right| \qquad \begin{array}{c} for \ a \ carnot \\ cycle \ based \ on \\ an \ ideal \ gas \end{array}$$
(14)

One suspects that if you do a lot of calculation involving integration, logarithms, and quantities like the specific heat ratio, and almost everything cancels to leave such a simple result as Equation 14, then there might be a deeper significance to the result than expected. Equation 14 was derived for a Carnot cycle operating with an ideal gas. It turns out that *the result is far more general* and has broad applications.

#### **Exercise 4**

A particular Carnot engine has an efficiency of 26.8%. That means that only 26.8% of  $Q_H$  comes out as useful work W and the rest, 73.2% is expelled at the low temperature  $T_L$ . The difference between the high and low temperature is 100 K ( $T_H - T_L = 100$  K). What are the values of  $T_H$  and  $T_L$ ? First express your answer in kelvins, then in degrees centigrade. (The answer should be familiar temperatures.)

#### **Exercise 5**

If you have a 100% efficient Carnot engine, what can you say about  $T_{\!H}$  and  $T_{\!L}?$ 

#### **Reversible Engines**

In our discussion of the principle of relativity, it was immediately clear why we developed the light pulse clock thought experiment. You could immediately see that moving clocks should run slow, and why that was a consequence of the principle of relativity.

We now have a new thought experiment, the Carnot engine, which is about as idealized as our light pulse clock. We have been able to calculate the efficiency of a Carnot engine, but it is not yet obvious what that has to do either with real engines, or more importantly with the second law of thermodynamics which we are studying. It is not obvious because we have not yet discussed one crucial feature of the Carnot engine.

The Carnot engine is explicitly designed to be reversible. As shown in Figure (14), we could start at point 1 and go to point 4 by an adiabatic expansion of the gas. During this expansion the gas would do work but no heat is allowed to flow in. Thus the work energy would come from thermal energy and the gas would cool from  $T_H$  to  $T_L$ .



**Figure 14** *The Carnot cycle run backward.* 

The next step of the reverse Carnot cycle is an isothermal expansion from a volume  $V_4$  to a volume  $V_3$ . During this expansion, the gas does an amount of work equal to the area under the curve as shown in Figure (15a). Since there is no change in the internal energy of a gas when the temperature of the gas remains constant, the heat flowing in equals the work done by the gas. This is the same amount of heat  $Q_L$  that flowed out when the engine ran forward.



a) During the isothermal expansion, some heat flows into the gas to supply the energy needed for the work done by the gas.



b) A lot more work is required, and a lot more heat is expelled when we compress the hot gas isothermally.

#### Figure 15

Heat flow when the Carnot cycle runs backward. Since more heat flows out than in, some work W is required for the cycle. The net effect is that the work W pumps heat out of the gas, giving us a refrigerator. In going from point 3 to point 2, we adiabatically compress the gas to heat it from the lower temperature  $T_L$  to the higher temperature  $T_H$ . Since the compression is adiabatic, no heat flows in or out.

In the final step from point 2 to point 1, we isothermally compress the gas back to its original volume  $V_1$ . Since the gas temperature remains constant at  $T_H$ , there is no change in thermal energy and all the work we do, shown as the area under the curve in Figure (15b), must be expelled in the form of heat flowing out of the cylinder. The amount of heat expelled is just  $Q_H$ , the amount that previously flowed in when the engine was run forward.

We have gone through the reverse cycle in detail to emphasize the fact that the engine should run equally well both ways. In the forward direction the engine takes in a larger amount of heat  $Q_H$  at the high temperature  $T_H$ , expels a smaller amount of heat  $Q_L$ at the lower temperature  $T_L$ , and produces an amount of useful work W equal to the difference  $Q_H - Q_L$ .

In the reverse process, the engine takes in a smaller amount of heat  $Q_L$  at the low temperature  $T_L$ , and expels a larger amount  $Q_H$  at the higher temperature  $T_H$ . Since more heat energy is expelled than taken in, an amount of work  $W = Q_H - Q_L$  must now be supplied to run the engine. When we have to supply work to pump out heat, we do not usually call the device an engine. The common name is a *refrigerator*. In a refrigerator, the refrigerator motor supplies the work W, a heat  $Q_L$  is sucked out of the freezer box, and a total amount of energy  $Q_L + W = Q_H$  is expelled into the higher room temperature of the kitchen. If we have a Carnot refrigerator running on an ideal gas, then the heats  $Q_L$  and  $Q_H$  are still given by Equation 14

$$\frac{Q_L}{Q_H} = \frac{T_L}{T_H}$$
(14epeated)

where  $T_L$  and  $T_H$  are the temperatures on a scale starting from absolute zero such as in the kelvin scale.

#### **Exercise** 6

How much work must a Carnot refrigerator do to remove 1000 joules of energy from its ice chest at 0° C and expel the heat into a kitchen at 27° C?

# **ENERGY FLOW DIAGRAMS**

Because of energy conservation, we can view the flow of energy in much the same way as the flow of some kind of a fluid. In particular we can construct flow diagrams for energy that look much like plumbing diagrams for water. Figure (16) is the energy flow diagram for a Carnot engine running forward. At the top and the bottom are what are called *thermal reservoirs*—large sources of heat at constant temperature (like swimming pools full of water). At the top is a thermal reservoir at the high temperature  $T_{H}$  (it could be kept at the high temperature by burning fuel) and at the bottom is a thermal reservoir at the low temperature  $T_{I}$ . For power plants, the low temperature reservoir is often the ocean or the cool water in a river. Or it may be the cooling towers like the ones pictured in photographs of the nuclear power plants at Three Mile Island.

In the energy flow diagram for the forward running Carnot engine, an amount of heat  $Q_H$  flows out of the high temperature reservoir, a smaller amount  $Q_L$  is expelled into the low temperature reservoir, and the difference comes out as useful work W. If the Carnot engine is run on an ideal gas,  $Q_H$  and  $Q_L$  are always related by  $Q_H/Q_L = T_H/T_L$ .



#### Figure 16

Energy flow diagram for a Carnot engine. Since energy is conserved, we can construct a flow diagram for energy that resembles a plumbing diagram for water. In a Carnot cycle,  $Q_H$  flows out of a "thermal reservoir" at a temperature  $T_H$ . Some of this energy goes out as useful work W and the rest,  $Q_L$ , flows into the low temperature thermal reservoir at a temperature  $T_L$ . Figure (17) is the energy flow diagram for a Carnot refrigerator. A heat  $Q_L$  is sucked out of the low temperature reservoir, an amount of work W is supplied (by some motor), and the total energy  $Q_H = Q_L + W$  is expelled into the high temperature reservoir.

#### **Maximally Efficient Engines**

We are now ready to relate our discussion of the Carnot cycles to the second law of thermodynamics. The statement of the second law we will use is that you cannot extract useful work from thermal energy at one temperature. (The colloquial statement of the first law of thermodynamics—conservation of energy—is that you can't get something for nothing. The second law says that you can't break even.)

Up until now we have had to point out that our formula for the efficiency of a Carnot engine was based on the assumption that we had an ideal gas in the cylinder. If we use the second law of thermodynamics, we can show that it is impossible to construct any engine, by any means, that is more efficient than the Carnot engine we have been discussing. This will be the main result of our thought experiment.



#### Figure 17

The Carnot refrigerator is a Carnot engine run backwards. The work W plus the heat  $Q_L$  equals the heat  $Q_H$  pumped up into the high temperature reservoir. Let us suppose that you have constructed a Super engine that takes in more heat  $Q_H^*$  from the high temperature reservoir, and does more work  $W^*$ , while rejecting the same amount of heat  $Q_L$  as a Carnot engine. In the comparison of the two engines in Figure (18) you can immediately see that your Super engine is more efficient than the Carnot engine because you get more work out for the same amount of heat lost to the low temperature reservoir.

Now let us run the Carnot engine backwards as a refrigerator as shown in Figure (19). The Carnot refrigerator requires an amount of work W to suck the heat  $Q_L$  out of the low temperature reservoir and expel the total energy  $Q_H = W + Q_L$  into the high temperature reservoir.

You do not have to look at Figure (19) too long before you see that you can use some of the work  $W^*$  that your Super engine produces to run the Carnot refrigerator. Since your engine is more efficient than the Carnot cycle,  $W^* > W$  and you have some work left over.



#### Figure 18

Comparison of the Super engine with the Carnot engine.





The next thing you notice is that you do not need the low temperature reservoir. All the heat expelled by your Super engine is taken in by the Carnot refrigerator. The low temperature reservoir can be replaced by a pipe and the new plumbing diagram for the combined Super engine and Carnot refrigerator is shown in Figure (20). The overall result of combining the super engine and Carnot refrigerator is that a net amount of work  $W_{net} = W^*$ -W is extracted from the high temperature reservoir. The net effect of this combination is to produce useful work from thermal energy at a single temperature, *which is a violation of the second law of thermodynamics*.

#### Exercise 7

Suppose you build a Super engine that takes the same amount of heat from the high temperature reservoir as a Carnot engine, but rejects less heat  $Q_L^* < Q_L$  than a Carnot engine into the low temperature. Using energy flow diagrams show what would happen if this Super engine were connected to a Carnot refrigerator. (You would still be getting useful work from thermal energy at some temperature. From what temperature reservoir would you be getting this work?)



#### Figure 20

If you connect the Super engine to the Carnot refrigerator, you can eliminate the low temperature reservoir and still get some work  $W_{net}$  out. This machine extracts work from a single temperature, in violation of the second law of thermodynamics.

# Reversibility

We have just derived the rather sweeping result that if the second law of thermodynamics is correct, you cannot construct an engine that is more efficient than a Carnot engine based on an ideal gas. You may wonder why the cycles based on an ideal gas are so special. It turns out that they are not special. What was special about the Carnot engine is that it was reversible, that it could be run backwards as a refrigerator. You can use precisely the same kind of arguments we just used to show that *all reversible engines must have precisely the same efficiency as a Carnot engine*. It is a requirement of the second law of thermodynamics.

There were two reasons we went through the detailed steps of constructing a Carnot engine using an ideal gas in a frictionless piston. The first was to provide one example of how an engine can be constructed. It is not a very practical example, commercial engines are based on different kinds of cycles. But the Carnot engine illustrates the basic features of all engines. In all engines the process must be repetitive, at least two temperatures must be involved, and only some of the heat extracted from the high temperature reservoir can be converted to useful work. Some heat must be expelled at a lower temperature.

While all reversible engines have the same efficiency, we have to work out at least one example to find out what that efficiency is. You might as well choose the simplest possible example, and the Carnot cycle using an ideal gas is about as simple as they get. Because of the second law of thermodynamics, you know that even though you are working out a very special example, the answer  $Q_H/Q_L = T_H/T_L$  applies to all reversible engines operating between two temperature reservoirs. This is quite a powerful result from the few pages of calculations in the appendix.

# APPLICATIONS OF THE SECOND LAW

During the oil embargo in the middle 1970s, there was a sudden appreciation of the consequences of the second law of thermodynamics, for it finally became clear that we had to use energy efficiently. Since that time there has been a growing awareness that there is a cost to producing energy that considerably exceeds what we pay for it. Burning oil and coal depletes natural limited resources and adds carbon dioxide to the atmosphere which may contribute to global climate changes. Nuclear reactors, which were so promising in the 1950s, pose unexpected safety problems, both now as in the example of Chernobyl, and in the very distant future when we try to deal with the storage of spent reactor parts. Hydroelectric power floods land that may have other important uses, and can damage the agricultural resources of an area as in the case of the Aswan Dam on the Nile River. More efficient use of energy from the sun is a promising idea, but technology has not evolved to the point where solar energy can supply much of our needs. What we have learned is that, for now, the first step is to use energy as efficiently as possible, and in doing this, the second law of thermodynamics has to be our guide.

During the 1950s and 60s, one of the buzz words for modern living was the *all electric house*. These houses were heated electrically, electric heaters being easy and inexpensive to install and convenient to use. And it also represents one of the most stupid ways possible to use energy. In terms of a heat cycle, it represents the 100% conversion of work energy into thermal energy, what we would have called in the last section, a 0% efficient engine. There are better ways of using electric power than converting it all into heat.

You can see where the waste of energy comes in when you think of the processes involved in producing electric power. In an electric power plant, the first step is to heat some liquid or gas to a high temperature by burning fuel. In a common type of coal or oil fired power plant, mercury vapor is heated to temperatures of 600 to 700 degrees centigrade. The mercury vapor is then used to run a mercury vapor turbine which cools the mercury vapor to around  $200^{\circ}$  C. This cooler mercury vapor then heats steam which goes through a steam turbine to a steam condenser at temperatures around  $100^{\circ}$  C. In a nuclear reactor, the first step is often to heat liquid sodium by having it flow through pipes that pass through the reactor. The hot sodium can then be used to heat mercury vapor which runs turbines similar to those in a coal fired plant. The turbines are attached to generators which produce the electric power.

Even though there are many stages, and dangerous and exotic materials used in power stations, we can estimate the maximum possible efficiency of a power plant simply by knowing the highest temperature  $T_H$  of the boiler, and the lowest temperature  $T_L$  of the condenser. If the power plant were a reversible cycle running between these two temperatures, it would take in an amount of heat  $Q_H$  at the high temperature and reject an amount of heat  $Q_L$  at the low temperature, where  $Q_H$  and  $Q_L$  are related by  $Q_H/Q_L = T_H/T_L$  (Eq. 14). The work we got out would be

$$W = Q_{\rm H} - Q_{\rm L} \qquad \begin{array}{l} amount \ of \ work \ from \\ a \ reversible \ cycle \end{array} \tag{15}$$

We would naturally define the efficiency of the cycle as the ratio of the work out to the heat energy in

efficiency = 
$$\frac{W}{Q_H} = \frac{Q_H - Q_L}{Q_H}$$
 (16)

If we solve Equation 14 for Q<sub>L</sub>

$$Q_{L} = Q_{H} \frac{T_{L}}{T_{H}}$$

and use this in Equation 16, we get

efficiency = 
$$\frac{Q_{H} - Q_{L}}{Q_{H}} = \frac{Q_{H} (1 - T_{L}/T_{H})}{Q_{H}}$$
  
efficiency =  $\frac{T_{H} - T_{L}}{T_{H}}$ 
efficiency of a reversible cycle (17)

Since by the second law of thermodynamics no process can be more efficient than a reversible cycle, Equation 17 represents the maximum possible efficiency of a power plant.

The important thing to remember about Equation 17 is that the temperatures  $T_H$  and  $T_L$  start from absolute zero. The only way we could get a completely efficient engine or power plant would be to have the low

temperature at absolute zero, which is not only impossible to achieve but even difficult to approach. You can see from this equation why many power plants are located on the shore of an ocean or on the bank of a large river. These bodies are capable of soaking up large quantities of heat at relatively low temperatures. If an ocean or river is not available, the power plant will have large cooling towers to condense steam. Condensing steam at atmospheric pressure provides a low temperature of T<sub>L</sub> = 100° C or 373 K.

Equation 15 also tells you why power plants run their boilers as hot as possible, using exotic substances like mercury vapor or liquid sodium. Here one of the limiting factors is how high a temperature turbine blades can handle without weakening. Temperatures as high as 450° C or around 720 K are about the limit of current technology. Thus we can estimate the maximum efficiency of power plants simply by knowing how high a temperature turbine blades can withstand, and that the plant uses water for cooling. You do not have to know the details of what kind of fuel is used, what kind of exotic materials are involved, or how turbines and electric generators work, as long as they are efficient. Using the numbers  $T_{\rm H} = 720 \, \rm k$ ,  $T_L = 373$  k we find that the maximum efficiency is about

$$\begin{array}{l} \underset{\text{efficiency}}{\text{maximum}} = \frac{T_{\text{H}} - T_{\text{L}}}{T_{\text{H}}} = \frac{720 - 373}{720} \\ = .48 \end{array} \tag{18}$$

Thus about 50% represents a theoretical upper limit to the efficiency of power plants using current technology. In practice, well designed power plants reach only about 33% efficiency due to small inefficiencies in the many steps involved.

You can now see why the all electric house was such a bad idea. An electric power plant consumes three times as much fuel energy as it produces electric energy. Then the electric heater in the all electric house turns this electric energy back into thermal energy. If the house had a modern oil furnace, somewhere in the order of 85% of the full energy can go into heating the house and hot water. This is far better than the 33% efficiency from heating directly by electricity.

#### **Electric Cars**

One of the hot items in the news these days is the electric car. It is often touted as the pollution free solution to our transportation problems. There are advantages to electric cars, but not as great an advantage as some new stories indicate.

When you plug in your electric car to charge batteries, you are not eliminating the pollution associated with producing useful energy. You are just moving the pollution from the car to the power plant, which may, however, be a good thing to do. A gasoline car may produce more harmful pollutants than a power station, and car pollutants tend to concentrate in places where people live creating smog in most major cities on the earth. (Some power plants also create obnoxious smog, like the coal fired plants near the Grand Canyon that are harming some of the most beautiful scenery in the country.)

In addition to moving and perhaps improving the nature of pollution, power plants have an additional advantage over car engines—they are more efficient. Car engines cannot handle as high a temperature as a power plant, and the temperature of the exhaust from a car is not as low as condensing steam or ocean water. Car engines seldom have an efficiency as high as 20%; in general, they are less than half as efficient as a power plant. Thus there will be a gain in efficiency in the use of fuel when electric cars come into more common use.

(One way electric cars have for increasing their efficiency is to replace brakes with generators. When going down a hill, instead of breaking and dissipating energy by heating the brake shoes, the gravitational potential energy being released is turned into electric energy by the generators attached to the wheels. This energy is then stored as chemical energy in the batteries as the batteries are recharged.)

#### The Heat Pump

There is an intelligent way to heat a house electrically, and that is by using a heat pump. The idea is to use the electric energy to pump heat from the colder outside temperature to the warmer inside temperature. Pumping heat from a cooler temperature to a warmer temperature is precisely what a refrigerator does, while taking heat from the freezer chest and exhausting it into the kitchen. The heat pump takes heat from the cooler outside and exhausts it into the house.

As we saw in our discussion, it takes work to pump heat from a cooler to a higher temperature. The ratio of the heat  $Q_L$  taken in at the low temperature, to the heat  $Q_H$ expelled at the higher temperature, is  $Q_H/Q_L = T_H/T_L$ for a maximally efficient refrigerator. The amount of work W required is  $W = Q_H - Q_L$ . The efficiency of this process is the ratio of the amount of heat delivered to the work required.

efficiency of  
heat pump = 
$$\frac{Q_H}{W} = \frac{Q_H}{Q_H - Q_L}$$
  
=  $\frac{T_H}{T_H - T_L}$  (19)

where the last step in Equation 19 used  $Q_L = Q_H T_L / T_H$ .

#### Exercise 8

Derive the last formula in Equation 19.

When the temperature difference  $T_H - T_L$  is small, we can get very high efficiencies, i.e., we can pump a lot of heat using little work. In the worst case, where  $T_L = 0$  and we are trying to suck heat from absolute zero, the efficiency of a heat pump is 1—heat delivered equals the work put in—and the heat pump is acting like a resistance heater.

To illustrate the use of a heat pump let us assume that it is freezing outside ( $T_L = 0^\circ C = 273 \text{ K}$ ) and you want the inside temperature to be  $27^\circ C = 300 \text{ K}$ . Then a heat pump could have an efficiency of

efficiency of heat  
pump running from 
$$= \frac{T_H}{T_H - T_L}$$
  
 $= \frac{300 \text{ k}}{300 \text{ k} - 273 \text{ k}}$   
 $= 11.1$ 

In other words, as far as the second law of thermodynamics is concerned, we should be able to pump eleven times as much heat into a house, when it is just freezing outside, as the amount of electrical energy required to pump the heat. Even if the electrical energy is produced at only 30% efficiency, we should still get  $.30 \times 11.1 = 3.3$  times as much heat into the house as by burning the fuel in the house at 100% conversion of fuel energy into heat.

#### **Exercise 9**

This so-called "heat of fusion" of water is 333kJ/kg. What that means is that when a kilogram (1 liter) of water freezes (going from 0° C water to 0° ice), 333 kilojoules of heat are released. Thus to freeze a liter of 0° C water in your refrigerator, the refrigerator motor has to pump  $333 \times 10^3$  joules of heat energy out of the refrigerator into the kitchen. The point of the problem is to estimate how powerful a refrigerator motor is required if you want to be able to freeze a liter of water in 10 minutes.

Assume that the heat is being removed at a temperature of 0°C and being expelled into a kitchen whose temperature is 30°C, and that the refrigerator equipment is 100% efficient. (We will account for a lack of efficiency at the end of this problem.)

In the United Sates, the power of motors is generally given in "horsepower", a familiar but archaic unit. The conversion factor is 1 horsepower = 746 watts, and a power of 1 watt is 1 joule per second.

Calculate the horsepower required, then double the answer to account for lack of efficiency. (Answer: 0.16 horsepower.)

#### Exercise 10

Here is a problem that should give you some practice with the concepts of efficiency. You have the choice of buying a furnace that converts heat energy of oil into heat in the house with 85% efficiency. I.e., 85% of the heat energy of the oil goes into the house, and 15% goes up the chimney. Or you can buy a heat pump which is half as efficient as a Carnot refrigerator. (This is a more realistic estimate of the current technology of refrigeration equipment.) At very low temperatures outside, heat pumps are not as efficient, and burning oil in your own furnace is more efficient. But if it does not get too cold outside, heat pumps are more efficient. At what outside temperature will the heat pump and the oil furnace have the same efficiency? Assume that the electric energy you use is produced by a power plant that is 30% efficient. (Answer: - 26° C.)

#### The Internal Combustion Engine

We finish this section on practical applications with a brief discussion of the internal combustion engine. The main point is to give an example of an engine that runs on a cycle that is different from a Carnot cycle. It is more difficult to apply the second law of thermodynamics to an internal combustion engine because it does not take heat in or expel heat at constant temperatures like the Carnot engine, but we can still analyze the work we get out using a pV diagram.

The pV diagram for an internal combustion engine is shown in Figure (21). At position 1, a fuel and air mixture have been compressed to a small volume  $V_1$ by the piston which is at the top of the cylinder. If it is a gasoline engine, the fuel air mixture is ignited by a spark from a sparkplug. If it is a diesel engine, the mixture of diesel fuel and air have been heated to the point of combustion by the adiabatic compression from point 4 to point 1 that has just taken place. One of the advantages of a diesel engine is that an electrical system to produce the spark is not needed. This is particularly important for boat engines where electric systems give all sorts of problems. (We said this was a section on practical applications.)

After ignition, the pressure and temperature of the gas rise rapidly to  $p_2$ ,  $T_2$  before the piston has had a chance to move. Thus the volume remains at  $V_1$  and the pV curve goes straight up to point 2. The heated gas then expands adiabatically, and cools some, driving the piston down to the bottom of the cylinder. This is the stroke from which we get work from the engine.

We now have a cylinder full of hot burned exhaust gases. In a 4 cycle engine, a valve at the top of the cylinder is opened, and a piston is allowed to rise, pushing the hot exhaust gases out into the exhaust pipes. Not much work is required to do this. This is the part of the cycle where (relatively) low temperature thermal energy is exhausted to the environment. While the piston goes back down, the valves are set so that a mixture of air and fuel are sucked into the piston. When the cylinder is at the bottom of the piston, we have a cool, low pressure fuel air mixture filling the full volume  $V_4$ . We are now at the position labeled (4) in Figure (21). It took two strokes (up and down) of the piston to go from position 3 to position 4.

In the final stroke, the valves are shut and the rising piston adiabatically compresses the gas back to the starting point  $p_1$ ,  $V_1$ ,  $T_1$ . While the increase in temperature during this compression is what is needed to ignite the diesel fuel, you do not want the temperature to rise enough to ignite the air gasoline mixture in a gasoline engine. This can sometimes happen in a gasoline engine, causing a knock in the engine, or sometimes allowing the engine to run for a while after you have shut off the ignition key and stopped the spark plug from functioning.





# ENTROPY

The second law of thermodynamics provided us with the remarkable result that the efficiency of all reversible engines is the same. Detailed calculation of this efficiency using a Carnot engine based on an ideal gas gave us a surprisingly simple formula for this efficiency, namely  $Q_H/Q_L = T_H/T_L$ . Our preceding examples involving car engines, power plants, refrigerators and heat pumps illustrate how important this simple relationship is to mankind.

When you do a calculation and a lot of stuff cancels out, it suggests that your result may have a simpler interpretation than you originally expected. This turns out to be true for our calculations of the heat flow in a Carnot engine. To get a new perspective on our equation for heat flow, let us write the equation in the form

$$\frac{Q_{\rm H}}{T_{\rm H}} = \frac{Q_{\rm L}}{T_{\rm L}} \tag{2}$$

In this form the equation for heat flow is beginning to look like a conservation law for the quantity Q/T. During the isothermal expansion, an amount  $Q_H/T_H$ of this quantity flowed into the piston. During the isothermal compression,  $Q_L/T_L$  flowed out. We find that if the engine is reversible, the amount of Q/T that flowed in is equal to the amount of Q/T that flowed out. The net effect is that there was no change in Q/T during the cycle.

To get a better insight into what this quantity Q/T may be, consider a nonreversible engine operating between  $T_H$  and  $T_L$ , an engine that would be less efficient than the Carnot engine. Assume that the less efficient than the Carnot engine both take in the same amount of heat  $Q_H$  at the high temperature  $T_H$ . Then the less efficient engine must do less work and expel more heat at the low temperature  $T_L$ . Thus  $Q_L$  for the less efficient engine is bigger than  $Q_L$  for the Carnot engine. Since  $Q_L/T_L = Q_H/T_H$  for the Carnot engine,  $Q_L/T_L$  must be greater than  $Q_H/T_H$  for the less efficient engine

$$\frac{Q_L}{T_L} > \frac{Q_H}{T_H} \qquad \begin{array}{c} \textit{for an engine that} \\ \textit{is less efficient than} \\ \textit{a Carnot engine} \end{array} \qquad (21)$$

We see that the inefficient engine expelled more Q/T than it took in. The inefficient, non reversible, engine creates Q/T while reversible engines do not.

As we have done throughout the course, whenever we encounter a quantity that is conserved, or sometimes conserved, we give it a name. We did this for linear momentum, angular momentum, and energy. Now we have a quantity Q/T that is unchanged by reversible engines, but created or increased by irreversible inefficient ones. We are on the verge of defining the quantity physicists call *entropy*. We say on the verge of defining entropy, because Q/T is not entropy itself; it represents the *change in entropy*. We can say that when the gas expanded, the entropy of the gas *increased* by  $Q_H/T_H$ . And when the gas was compressed, the entropy decreased by  $Q_L/T_L$ .

For a reversible engine there is no net change in entropy as we go around the cycle. But for an irreversible, inefficient engine, more entropy comes out than goes in during each cycle. The net effect of an inefficient engine is to create entropy.

What is this thing called entropy that is created by inefficient irreversible engine? Consider the most inefficient process we can imagine—the electric heater which converts useful work in the form of electrical energy into heat. From one point of view, the device does nothing but create entropy. If the heater is at a temperature T, and the electric power into the heater is W watts, all this energy is converted to heat and entropy is produced at a rate of W/T in units of entropy per second. (Surprisingly, there is no standard name for a unit of entropy. The units of Q/T are of course joules/ kelvin, the same as Boltzman's constant.)

The process of converting energy in the form of useful work into the random thermal energy of molecules can be viewed as the process of turning order into disorder. Creating entropy seems to be related to creating disorder. But the surprising thing is that we have an explicit formula Q/T for changes in entropy. How could it be possible to measure disorder, to have an explicit formula for changes in disorder? This question baffled physicists for many generations. Ludwig Boltzman proposed that entropy was related to the number of ways that a system could be arranged. Suppose, for example, you go into a woodworking shop and there are a lot of nails on the wall with tools hanging on them. In one particular woodworking shop you find that the carpenter has drawn an outline of the tool on the wall behind the nail. You enter her shop you find that the tool hanging from each nail exactly matches the outline behind it. Here we have perfect order, every tool is in its place and there is one and only one way the tools can be arranged. We would say that, as far as locating tools is concerned the shop is in perfect order, it has no disorder or entropy.

On closer inspection, we find that the carpenter has two saws with identical outlines, a crosscut saw and a rip saw. We also see that the nails are numbered, and see the cross cut saw on nail 23 and the rip saw on nail 24. A week later when we come back, the cross cut is on nail 24 and the rip saw on 23. Thus we find that her system is not completely orderly, for there are two different ways the saws can be placed. This way of organizing tools has some entropy.

A month later we come back to the shop and find that another carpenter has taken over and painted the walls. We find that there are still 25 nails and 25 tools, but now there is no way to tell which tool belongs on which nail. Now there are many, many ways to hang up the tools and the system is quite disordered. We have the feeling that this organization, or lack of organization, of the tools has quite high entropy.

To put a numerical value on how disorganized the carpenter shop is, we go to a mathematician who tells us that there are N! ways to hang N tools on N nails. Thus there are  $25! = 1.55 \times 10^{25}$  different ways the 25 tools can be hung on the 25 nails. We could use this number as a measure of the disorder of the system, but the number is very large and increases very rapidly with the number of tools. If, for example, there were 50 tools hung on 50 nails, there would be  $3.04 \times 10^{64}$  different ways of hanging them. Such large numbers are not convenient to work with.

When working with large numbers, it is easier to deal with the logarithm of the number than the number itself. There are approximately  $10^{51}$  protons in the earth. The log to the base 10 of this number is 51 and the natural logarithm,  $\ln(10^{51})$ , is 2.3 times bigger or 117. In discussing the number of protons in the earth, the number 117 is easier to work with than  $10^{51}$ , particularly if you have to write out all the zeros.

If we describe the disorder of our tool hanging system in terms of the logarithm of the number of ways the tools can be hung, we get a much more reasonable set of numbers, as shown in Table 1.

Setup	Number of ways to arrange tools	Logarithm of the number of ways		
all tools have unique positions	s 1	0		
two saws can be interchanged	2	.7		
25 tools on unmarked nails	$1.5 \times 10^{25}$	58		
50 tools on unmarked nails	$3.0 \times 10^{65}$	148		

#### TABLE 1

The table starts off well. If there is a unique arrangement of the tools, only one way to arrange them, the logarithm of the number of ways is 0. This is consistent with our idea that there is no disorder. As the number of tools on unmarked nails increases, the number of ways they can be arranged increases at an enormous pace, but the logarithm increases at a reasonable rate, approximately as fast as the number of tools and nails. This logarithm provides a reasonable measure of the disorder of the system. We could define the entropy of the tool hanging system as the logarithm of the number of ways the tools could be hung. One problem, however, is that this definition of entropy would have different dimensions than the definition introduced earlier in our discussions of engines. There changes in entropy, for example  $Q_H/T_H$ , had dimensions of joules/kelvin, while our logarithm is dimensionless. However, this problem could be fixed by multiplying our dimensionless logarithm by some fundamental constant that has the dimensions of joules/ kelvin. That constant, of course, is Boltzman's constant k, where  $k = 1.38 \times 10^{-23}$  joules/kelvin. We could therefore take as the formula for the entropy (call it S) of our tool hanging system as

$$S = k \ln(n) \qquad \begin{array}{c} entropy of our tool \\ hanging system \end{array} \tag{2}$$

where n is the number of ways the tools can be hung. Multiplying our logarithm by k gives us the correct dimensions, but very small values when applied to as few items as 25 or 50 tools.

Equation 22 appears on Boltzman's tombstone as a memorial tohis main accomplishment in life. Boltzman believed that Equation 22 should be true in general. That, for example, it should apply to the atoms of the gas inside the cylinder of our heat engine. When heat flows into the cylinder and the entropy increases by an amount  $Q_H/T_H$ , the number of ways that the atoms could be arranged should also increase, by an amount we can easily calculate using Equation 22. Explicitly, if before the heat flowed in there were  $n_{old}$  ways the atoms could be arranged, and after the heat flowed in  $n_{new}$  ways, then Equation 22 gives

change in  
entropy = 
$$k \ln (n_{new}) - k \ln (n_{old}) = \frac{Q_H}{T_H}$$

Since  $\ln(n_{new}) - \ln(n_{old}) = \ln(n_{new}/n_{old})$ , we get

$$\ln\left(\frac{n_{\text{new}}}{n_{\text{old}}}\right) = \frac{Q_{\text{H}}}{kT_{\text{H}}} \tag{23}$$

Taking the exponent of both sides of Equation 16, using the fact that  $e^{\ln (x)} = x$ , we get

$$\frac{n_{\text{new}}}{n_{\text{old}}} = e^{Q_{\text{H}}/kT_{\text{H}}}$$
<sup>(24)</sup>

Thus Boltzman's equation gives us an explicit formula for the fractional increase in the number of ways the atoms in the gas atoms in the cylinder can be arranged.

Boltzman committed suicide in 1906, despondent over the lack of acceptance of his work on the statistical theory of matter, of which Equation 22 is the cornerstone. And in 1906 it is not too surprising that physicists would have difficulty dealing with Boltzman's equation. What is the meaning of the number of ways you can arrange gas atoms in a cylinder? From a Newtonian perspective, there are an infinite number of ways to place just one atom in a cylinder. You can count them by moving the atoms an infinitesimal distance in any direction. So how could it be that  $10^{24}$ atoms in a cylinder have only a finite way in which they can be arranged?

This question could not be satisfactorily answered in 1906, the answer did not come until 1925 with the discovery of quantum mechanics. In a quantum picture, an atom in a cylinder has only certain energy levels, an idea we will discuss later in Chapter 35. Even when you have  $10^{24}$  atoms in the cylinder, the whole system has only certain allowed energy levels. At low temperatures the gas does not have enough thermal energy to occupy very many of the levels. As a result the number of ways the atoms can be arranged is limited and the entropy is low. As the temperature is increased, the gas atoms can occupy more levels, can be arranged in a greater number of ways, and therefore have a greater entropy.

The concept of entropy provides a new definition of absolute zero. A system of particles is at absolute zero when it has zero entropy, when it has one uniquely defined state. We mentioned earlier that quantum mechanics requires that a confined particle has some kinetic energy. All the kinetic energy cannot be removed by cooling. This gives rise to the so-called *zero point energy* that keeps helium a liquid even at absolute zero. However, a bucket of liquid helium can be at absolute zero as long as it is in a single unique quantum state, even though the atoms have zero point kinetic energy. In our discussion of temperature in the last chapter, we used the ideal gas thermometer for our experimental definition of temperature. We pointed out, however, that this definition would begin to fail as we approached very low temperatures near absolute zero. At these temperatures we need a new definition which agrees with the ideal gas thermometer definition at higher temperatures. The new definition which is used by the physics community, is based on the efficiency of a reversible engine or heat cycle. You can measure the ratio of two temperatures T<sub>H</sub> and T<sub>C</sub>, by measuring the heats  $Q_H$  and  $Q_C$  that enter and leave the cycle, and use the formula  $T_H/T_C = Q_H/Q_C$ . Since this formula is based on the idea that a reversible cycle creates no entropy  $(Q_H/T_H = Q_C/T_C)$ , we can see that the concept of entropy forms the basis for the definition of temperature.

#### **The Direction of Time**

We began the chapter with a discussion of a demonstration that looked funny. We set the strobe so that the water drops appeared to rise from the plate in the bucket and enter the hose. Before our discussion of the second law of thermodynamics, we could not find any law of physics that this backward process appeared to violate. Now we can see that the launching of the drops from the plate is a direct contradiction of the second law. In that process, heat energy in the bucket converts itself at one temperature into pure useful work that launches the drop.

When you run a moving picture of some action backwards, effectively reversing the direction of time, in most cases the only law of physics that is violated is the second law of thermodynamics. The only thing that appears to go wrong is that disordered systems appear to organize themselves on their own. Scrambled eggs turn into an egg with a whole yoke just by the flick of a fork. Divers pop out of swimming pools propelled, like the drops in our demonstration, by the heat energy in the pool (see movie). All these funny looking things require remarkable coincidences which in real life do not happen. For a while there was a debate among physicists as to whether the second law of thermodynamics was the only law in nature that could be used to distinguish between time running forward and time running backward.

When you study processes like the decay of one kind of elementary particle into another, the situation is so simple that the concepts of entropy and the second law of thermodynamics do not enter into the analysis. In these cases you can truly study whether nature is symmetric with the respect to the reversal of time. If you take a moving picture of a particle decay, and run the movie backwards, will you see a process that can actually happen? For example if a muon decays into an electron and a neutrino, as happened in our muon lifetime experiment, running that moving picture backward would have neutrinos coming in, colliding with an electron, creating a muon. Thus, if the basic laws of physics are truly symmetric to the reversal of time, it should be possible for a neutrino and an electron to collide and create a muon. This process is observed.

In 1964 Val Fitch and James Cronin discovered an elementary particle process which indicated that nature was not symmetric in time. Fitch found a violation of this symmetry in the decay of a particle called the neutral k meson. For this discovery, Fitch was awarded the Nobel prize in 1980. Since the so-called "weak interaction" is responsible for the decay of k mesons, the weak interaction is not fully symmetric to the reversal of time. The second law of thermodynamics is not the only law of physics that knows which way time goes.



**Movie** Time reversed motion picture of dive

# APPENDIX: CALCULATION OF THE **EFFICIENCY OF A CARNOT CYCLE**

The second law of thermodynamics tells us that the efficiency of all reversible heat engines is the same. Thus if we can calculate the efficiency of any one engine, we have the results for all. Since we have based so much of our discussion on the Carnot engine running on an ideal gas, we will calculate the efficiency of that engine.

To calculate the efficiency of the ideal gas Carnot engine, we need to calculate the amount of work we get out of (or put into) isothermal and adiabatic expansions. With these results, we can then calculate the net amount of work we get out of one cycle and then the efficiency of the engine. To simplify the formulas, we will assume that our engine is running on one mole of an ideal gas.

# **Isothermal Expansion**

Suppose we have a gas at an initial volume  $V_1$ , pressure p<sub>1</sub>, temperature T, and expand it isothermally to a volume  $V_2$ , pressure  $p_2$ , and of course the same temperature T.

The P-V diagram for the process is shown in Figure (A-1). The curve is determined by the ideal gas law, which for 1 mole of an ideal gas is

The work we get out of the expansion is the shaded area under the curve, which is the integral of the pressure curve from  $V_1$  to  $V_2$ .



Isothermal expansion.

Using Equation 1, we get

$$W = \int_{V_1}^{V_2} p dV = \int_{V_1}^{V_2} \frac{RT}{V} dV$$
$$= RT \int_{V_1}^{V_2} \frac{dV}{V} = RT \ln V \Big|_{V_1}^{V_2}$$
$$= RT \Big[ \ln (V_2) - \ln (V_1) \Big]$$
$$= RT \ln \Big( \frac{V_2}{V_1} \Big)$$
(A-2)

Thus the work we get out is RT times the logarithm of the ratio of the volumes.

# Adiabatic Expansion

It is a bit trickier to calculate the amount of work we get out of an adiabatic expansion. If we start with a mole of ideal gas at a volume  $V_1$ , pressure  $p_1$ , and temperature  $T_{\rm H}$ , the gas will cool as it expands because the gas does work and we are not letting any heat in. Thus when the gas gets to the volume  $V_2$ , at a pressure  $p_2$ , its temperature T<sub>C</sub> will be cooler than its initial temperature T<sub>H</sub>.

The pV diagram for the adiabatic expansion is shown in Figure (A-2). To get an equation for the adiabatic expansion curve shown, let us assume that we change the volume of the gas by an infinitesimal amount  $\Delta V$ . With this volume change, there will be an infinitesimal pressure drop  $\Delta p$ , and an infinitesimal temperature



Adiabatic expansion.

drop  $\Delta T$ . We can find the relationship between these small changes by differentiating the ideal gas equations. Starting with

pv=RT

and differentiating we get

$$p\Delta V + (\Delta p)V = R\Delta T \tag{A-3}$$

We now have to introduce the idea that the expansion is taking place adiabatically, i.e., that no heat is entering. That means that the work  $p\Delta V$  done by the gas during the infinitesimal expansion  $\Delta V$  must all have come from thermal energy. But the decrease in thermal energy is  $C_V\Delta T$ . Thus we have from conservation of energy

$$p\Delta V + C_V \Delta T = 0$$

or

$$\Delta T = \frac{-p\Delta V}{C_V} \tag{A-4}$$

The - (minus) sign tells us that the temperature drops as work energy is removed.

Using Equation A-4 for  $\Delta T$  in Equation A-3 gives

$$p\Delta V + \Delta pV = R\left(\frac{-p\Delta V}{C_V}\right)$$

Combining the  $p\Delta V$  terms gives

7

$$p\Delta V \left(1 + \frac{R}{C_V}\right) + \Delta pV = 0$$

$$p\Delta V \left(\frac{C_V + R}{C_V}\right) + \Delta pV = 0 \quad (A-5)$$

Earlier in the chapter, in Equation 9, we found that for an ideal gas,  $C_V$  and  $C_p$  were related by  $C_p = C_V + R$ . Thus Equation A-5 simplifies to

$$p\Delta V \left(\frac{C_p}{C_V}\right) + \Delta pV = 0 \tag{A-6}$$

It is standard notation to define the ratio of specific heats by the constant  $\gamma$ 

$$\frac{C_{p}}{C_{V}} \equiv \gamma \tag{A-7}$$

thus Equation A-6 can be written in the more compact form

$$\gamma p \Delta V + \Delta p V = 0 \tag{A-8}$$

The next few steps will look like they were extracted from a calculus text. They may or may not be too familiar, but you should be able to follow them step-bystep.

First we will replace  $\Delta V$  and  $\Delta p$  by dV and dp to indicate that we are working with calculus differentials. Then dividing through by the product pV gives

$$\gamma \frac{\mathrm{dV}}{\mathrm{V}} + \frac{\mathrm{dp}}{\mathrm{p}} = 0 \tag{A-9}$$

Doing an indefinite integration of this equation gives

$$\gamma \ln (V) + \ln (p) = \text{const}$$
 (A-10)

The  $\gamma$  can be taken inside the logarithm to give

$$\ln(V^{\gamma}) + \ln(p) = \text{const}$$
 (A-11)

Next exponentiate both sides of Equation A-11 to get

 $e^{\ln(V^{\gamma}) + \ln(p)} = e^{const} = another const$  (A-12)

where  $e^{\text{const}}$  is itself a constant. Now use the fact that  $e^{a+b} = e^a e^b$  to get

$$e^{\ln(\mathbf{V}^{\gamma})}e^{\ln(\mathbf{p})} = \text{const} \tag{A-13}$$

Finally use  $e^{\ln(x)} = x$  to get the final result

$$pV^{\gamma} = const$$
  $adiabatic expansion$  (A-14)

Equation A-14 is the formula for the adiabatic curve seen in Figure (A-2). During an isothermal expansion, we have pV=RT where T is a constant. Thus if we compare the formulas for isothermal and adiabatic expansions, we have for any ideal gas

$$pV = const \quad isothermal expansion$$
  

$$pV^{\gamma} = const \quad adiabatic expansion$$
  

$$\gamma = C_p / C_V \quad ratio \ of \ specific \ heats$$
  

$$C_p = C_V + R$$
  
(A-15)

# **The Carnot Cycle**

We now have the pieces in place to calculate the efficiency of a Carnot cycle running on one mole of an ideal gas. The cycle is shown in Figure (11) repeated here as Figure (A-3).

During the isothermal expansion from point 1 to 2, the amount of heat that flows into our mole of gas is equal to the work one by the gas. By Equation A-2, this work is

$$W_{12} = Q_H = RT_H \ln (V_2/V_1)$$
 (A-16)

pressure



The Carnot cycle.

The heat  $Q_L$  expelled at the low temperature  $T_L$  is equal to the work we do compressing the gas isothermally in going from point 3 to 4. This work is

$$W_{34} = Q_L = RT_C \ln(V_4/V_3)$$
 (A-17)

Taking the ratio of Equations A-16 to A-17 we get

$$\frac{Q_{\rm H}}{Q_{\rm L}} = \frac{T_{\rm H} \ln \left( V_2 / V_1 \right)}{T_{\rm C} \ln \left( V_4 / V_3 \right)} \tag{A-18}$$

The next step is to calculate the ratio of the logarithms of the volumes using the adiabatic expansion formula  $pV^{\gamma} = constant$ .

In going adiabatically from 2 to 3 we have

$$p_2 V_2^{\gamma} = p_3 V_3^{\gamma}$$
 (A-19)

and in going from 4 to 1 adiabatically we have

$$p_4 V_4^{\gamma} = p_1 V_1^{\gamma} \tag{A-20}$$

Finally, use the ideal gas law pV = RT to express the pressure p in terms of volume and temperature in Equations A-19 and A-20. Explicitly use

$$p_{1} = RT_{H} / V_{1}; \quad p_{2} = RT_{H} / V_{2}$$
  

$$p_{3} = RT_{H} / V_{3}; \quad p_{4} = RT_{H} / V_{4}$$
(A-21)

to get for Equation (19)

$$\frac{\mathrm{RT}_{\mathrm{H}}}{\mathrm{V}_{2}} \mathrm{V}_{2}^{\gamma} = \frac{\mathrm{RT}_{\mathrm{C}}}{\mathrm{V}_{3}} \mathrm{V}_{3}^{\gamma}$$

or

$$T_{\rm H} V_2^{\gamma - 1} = T_{\rm C} V_3^{\gamma - 1}$$
 (A-22)

and similarly for Equation (20)

$$T_{\rm H} V_1^{\gamma - 1} = T_{\rm C} V_4^{\gamma - 1} \tag{A-23}$$

as you can check for yourself.

If we divide Equation A-22 by A-23 the temperatures  $T_{\rm H}$  and  $T_{\rm C}$  cancel, and we get

$$\frac{V_2^{\gamma - 1}}{V_1^{\gamma - 1}} = \frac{V_3^{\gamma - 1}}{V_4^{\gamma - 1}}$$

or

$$\left(\frac{\mathbf{V}_2}{\mathbf{V}_1}\right)^{\gamma-1} = \left(\frac{\mathbf{V}_3}{\mathbf{V}_4}\right)^{\gamma-1} \tag{A-24}$$

Taking the  $(\gamma - 1)$ th root of both sides of Equation A-24 gives simply

$$\left(\frac{\mathbf{V}_2}{\mathbf{V}_1}\right) = \left(\frac{\mathbf{V}_3}{\mathbf{V}_4}\right) \tag{A-25}$$

Since  $V_2 / V_1 = V_3 / V_4$ , the logarithms in Equation A-18 cancel, and we are left with the surprisingly simple result

 $\frac{Q_H}{Q_L} = \frac{T_H}{T_L}$  (14epended)

which is our Equation 14 for the efficiency of a Carnot cycle. As we mentioned, when you are doing a calculation and a lot of stuff cancels to give a simple result, there is a chance that your result is more general, or has more significance than you expected. In this case, Equation 14 is the formula for the efficiency of any reversible engine, no matter how it is constructed. We happened to get at this formula by calculating the efficiency of a Carnot engine running on one mole of an ideal gas.

# Chapter 19 The Electric Interaction Atomic & Molecular Forces

# THE FOUR BASIC INTERACTIONS

The world around us is a complex place with enormous variety of a myriad of interactions. But if you look in the right places, from the right point of view, you may find great simplicity. Planetary motion is one example. If you look at the sun and planets alone, ignoring things on a larger scale like other stars and galaxies, and on a smaller scale like the makeup of the planets and the atmosphere of the sun, you have a system of 10 objects whose behavior is accurately determined by a single force law. The system is simple enough that mankind learned about physics by studying it.

In this and the next chapter we take a first look at several of the basic patterns and laws of physics. Perhaps the most important discovery in the twentieth century, actually more of a gradual realization made during the first half of the twentieth century, was that all of the phenomena of nature, everything we see around us, can be explained in terms of **four basic forces or interactions**. In some circumstances, as in the case of planetary motion, a single force dominates and the structures it creates are obvious. In most cases however, the structures are complex and it is no easy task to uncover the underlying forces.

What we will do in these two chapters is to describe the four basic interactions, focusing our attention on examples where the action of the force is most clearly seen. For the **gravitational force**, a planetary system provides the most clear and detailed example of a structure created by gravity. Using this example as a guide we can see that larger structures like globular clusters and galaxies are an extension of the gravitational interaction to more complex situations. Although we do not attempt to calculate in detail the motion of the millions or billions of stars in such objects, we gain an intuitive feeling for the kind of structures the gravitational interaction creates.

The *electric interaction*, the second of the four basic interactions to be discovered, is most clearly seen at an atomic level. To introduce the electric interaction, we will take the simple point of view that atoms consist of a tiny nucleus made from protons and neutrons, surrounded by electrons held to the nucleus by the electric force. The model should be familiar, it is essentially a scaled down version of the solar system, with an electric force replacing the gravitational force. After a brief discussion of the kinds of nuclei that can be made from protons and neutrons, we will look at the properties of the electric interaction that give rise to complete atoms. This involves the concept of electric charge, the fact that electricity, like gravity, is a  $1/r^2$  force law, and the fact that, on an atomic scale, the electric force is much, much stronger than gravity.



Figure 0

The classical model of the hydrogen atom closely resembles a planetary system

When you bring complete atoms close to each other, there are weak residual electric forces which result from small distortions of the atomic structure. These weak residual forces are the **molecular forces** that hold atoms together to form molecules, crystals and most of the variety we see in the world about us. In this chapter we will consider only the simplest example of a molecular force to see how such residual forces arise when atoms interact.

In the next chapter we leave the scale of atoms and molecules and look down inside the atomic nucleus, where we find two more forces at work. There is an attractive force, called the **nuclear interaction**, that is even stronger than electricity. It holds the nucleus together against the electric repulsion between the proton in the nucleus. There is also another force, called the **weak interaction**, which allows neutrons to decay into protons and protons into neutrons (the  $\beta$ decay reaction discussed in Chapter 6). The nature of nuclear reactions, the stability of atomic nuclei, and the abundance of the elements depend upon a delicate interplay of the nuclear, the weak, and the electric interactions.

In the past 30 years we have been able to sharpen our view of nuclear and subnuclear matter. We have been able to look inside the proton and neutron and discover that they are not elementary particles, but instead, composite objects made from **quarks**. We have also found that the basic nuclear force is the one between quarks. The force that holds protons together in the nucleus is a residual of the quark force, just as molecular forces that hold atoms together are a residual of the electric force. Because of a mathematical analogy to the theory of color, the force between quarks is called the **color force**. The color force now replaces the residual nuclear force as one of the four basic interactions.

Once we see how nature can be explained in terms of just four basic forces, we cannot help wondering why the number is four. Are there more basic forces, some of them yet undetected? Or are there fewer than four basic forces, some of the four being equivalent on a more fundamental level? Einstein spent the latter half of his life trying to find a unified way of viewing the gravitational and electric interactions. He tried to find a single more fundamental theory from which both electricity and gravity emerged. He did not succeed, and no one else has yet done so either. However from the knowledge gained by looking inside the proton and neutron, the knowledge that led to the discovery of quarks, Steven Weinberg and Sheldon Glashow were able to construct a theory that unified the electric and weak interactions. These two forces which appeared so different in their effects on matter, turn out to be two components of a more fundamental interactions. Thus we are now down to three basic interactions. Why three? Can these be unified? We do not know yet.

We end this discussion with another look at the gravitational interaction. On an atomic scale, gravity is so weak compared to electricity that only recently has it been experimentally determined that electrons fall down rather than up in the earth's gravitational field. The only reason we human beings personally know about gravity is the fact that we are standing on a huge chunk of matter, the earth. It takes a lot of matter to create a big enough gravitational force for us to notice.

But there is a lot of matter in the universe. Sometimes, as in the case of a neutron star, so much matter is packed in such a small space that the gravitational force becomes stronger than the electric force. In a neutron star, gravity has forced the electrons back into the nucleus, to form pure nuclear matter. If the neutron star gets too big, if gravity gets a bit stronger, it can overwhelm the nuclear force and crush the star to form a black hole. Gravity, a force so weak that we could barely detect it using the Cavandish experiment, can become the strongest of all forces.

# **ATOMIC STRUCTURE**

To set the stage for our discussion of the electric interaction, we will first construct a brief overview of atomic and nuclear structure. The components of atoms and nuclei that are of interest are the proton and neutron found in the nucleus, and the electron which orbits outside. Protons and neutrons are each about 1836 times as massive as the electron, thus most of the mass of an atom is located in the nucleus.

The simplest of all atoms is hydrogen with one proton for a nucleus and one electron outside. The electron is attracted to the proton by a  $1/r^2$  electric force, just as the earth is attracted to the sun by the  $1/r^2$  gravitational force. And because the proton is much more massive than the electron, the proton sits nearly at rest at the center of the atom while the electron orbits outside, much as the earth orbits the sun.

Also like the solar system, the atom is mostly empty space. A proton has a diameter of about  $10^{-13}$  cm, while a hydrogen atom is one hundred thousand times bigger. If the hydrogen atom were enlarged to the point where the proton nucleus were the size of the sun, the electron would be orbiting out at a distance over 10 times the radius of the pluto's orbit. In this sense there is more empty space in an atom than in the solar system.

In Newton's law of gravity, the gravitational force on an object is proportional to an object's mass. The fact that all gravitational forces are attractive can be viewed as a consequence of the fact that there is only positive mass. In the electric interaction, there are both attractive and repulsive electric forces. We will see that for the electric interaction, the concept of electric charge plays a role similar to that of mass for the gravitational interaction. The existence of both attractive and repulsive electric forces leads to having both positive and negative charge. One new feature of having both attractive and repulsive electric forces is that the net electric force between two objects can be zero, due to the cancellation of attractive and repulsive components. This cancellation of electric force can be represented by a cancellation of electric charge, giving us an object which we say is *electrically neutral*. Because there are no repulsive gravitational forces and no negative mass, there is no such thing as a gravitationally neutral object.

When an atom has the same number of electrons in orbit as the number of protons in its nucleus, the atom is electrically neutral. If you have two electrically neutral atoms separated by a reasonable distance, like the atoms in a gas, the electric forces between the atoms cancel. Thus neutral atoms in a gas move by each other with almost no interaction. There is an interaction only when the atoms get too close in a collision and the electric forces no longer cancel.

Atoms are classified by the number of protons in the nucleus. If the nucleus has one proton, the atom belongs to the element hydrogen. If there are 2 protons in the nucleus, it is a helium atom. Three protons gives us lithium, on up through the periodic table. The largest naturally occurring atom, on the earth at least, is uranium with 92 protons. Atoms with over 100 protons in the nucleus have been created artificially.

The periodic table, and the classification of the elements, were developed by chemists studying the chemical properties of matter. However chemical reactions, with the possible exception of cold fusion, have virtually no effect on the atomic nucleus. If you have a lead nucleus with its 82 protons, there is no set of chemical reactions that can change it to a gold nucleus with 79 protons. This is what doomed the alchemists of the middle ages to failure.

The chemistry of an atom depends upon the behavior of the electrons in an atom, and the electron behavior depends significantly on the number of electrons. Since an electrically neutral atom has the same number of electrons as protons, different elements with different numbers of protons have different numbers of electrons and thus different chemical properties. For example, hydrogen with one electron is an excellent fuel, helium with 2 electrons is chemically inert, and lithium with 3 electrons is a highly reactive alkali metal.

The periodic table is not merely a list of atoms according to the number of protons in the nucleus. The table exhibits many striking patterns or regularities in the chemical behavior of the elements. For example, helium with 2 electrons, neon with 10, argon with 18, krypton with 36, xenon with 54 and radon with 86 electrons are all chemically inert gases. These so-called noble gases enter into few if any chemical reactions. Now add one electron to each of these atoms (and one proton to the nucleus), and you get lithium (3 electrons), sodium (11 electrons), potassium (19 electrons), etc., all reactive alkali metals.

The patterns in the chemical properties of the elements exhibited by the periodic table are a consequence of the electric interaction, but they cannot be explained using Newtonian mechanics. Scientists had to wait until the discovery of quantum mechanics before a detailed explanation of the periodic table unfolded. After we have discussed some of the basic ideas of quantum mechanics in later chapters, we will see that there are fairly simple explanations for the main features of the periodic table, like the difference between noble gasses and alkali metals mentioned above. For now, however, where we have only developed a background of Newtonian mechanics, we will go no further than treating the periodic table as a list of the elements according to the number of protons in the nucleus.

Element	Chemical symbol	No. of protons	No. of neutrons	Element	Chemical symbol	No. of protons	No. of neutrons
Hydrogen	Н	1	0	Iodine	Ι	53	74
Helium	He	2	2	Xenon	Xe	54	78
Lithium	Li	3	4	Cesium	Cs	55	78
Beryllium	Be	4	5	Barium	Ba	56	82
Boron	В	5	6	Lanthanum	La	57	82
Carbon	С	6	6	Cerium	Ce	58	82
Nitrogen	Ν	7	7	Praseodymium	ı Pr	59	82
Oxygen	О	8	8	Neodymium	Nd	60	82
Fluorine	F	9	10	Promethium	Pm	61	86
Neon	Ne	10	10	Samarium	Sm	62	90
Sodium	Na	11	12	Europium	Eu	63	90
Magnesium	Mg	12	12	Gadolinium	Gd	64	94
Aluminum	Al	13	14	Terbium	Th	65	94
Silicon	Si	14	14	Dysprosium	Dv	66	98
Phosphorus	Р	15	16	Holmium	Ho	67	98
Sulfur	S	16	16	Erbium	Er	68	98
Chlorine	Cl	17	18	Thulium	Tm	69	100
Argon	A	18	22	Vtterbium	Thi Vh	70	100
Potassium	K	19	20	Lutetium	IU Iu	70	104
Calcium	Ca	20	20	Luctum	Lu Uf	71	104
Scandium	Sc	21	20	Tannuni	111 Te	72	108
Titanium	Ti	21	26	T antaium Tun astan	1a W	75	108
Vanadium	V	22	28	Dhanium	W Da	74	110
Chromium	Čr.	23	28	Cominan	Re	75	112
Manganese	Mn	25	30	Usmium	Us	70	110
Iron	Fe	25	30	Iridium	lr D	77	110
Cohalt		20	30	Platinum	Pt	/8	117
Violal		27	32	Gold	Au	79	122
Common		20	30	Mercury	Hg	80	122
Copper	Cu Zu	29	54 24	Thallium	TI	81	124
	Zn	30 21	34 29	Lead	Pb	82	126
Gallium	Ga	31	38	Bismuth	Bi	83	126
Germanium	Ge	32	42	Polonium	Ро	84	124 (3 yr)
Arsenic	As	33	42	Astatine	At	85	125 (8 hr)
Selenium	Se	34	46	Radon	Rn	86	136 (3 days)
Bromide	Br	35	44	Francium	Fr	87	136 (21 min)
Krypton	Kr	36	48	Radium	Ra	88	138 (1622 yr)
Rubidium	Rb	37	48	Actinium	Ac	89	138 (22 hr)
Strontium	Sr	38	50	Thorium	Th	90	140 (80,000 yr)
Yttrium	Y	39	50	Protactinium	Pa	91	140 (34,000 yr)
Zirconium	Zr	40	50	Uranium	U	92	146 (4.5 billion yr)
Niobium	Nb	41	52	Neptunium	Np	93	144 (2.2 million yr)
Molybdenum	Mo	42	56	Plutonium	Pu	94	145 (24,000 yr)
Technetium	Tc	43	54 (> 100 yr)	Americium	Am	95	144 (490 yr)
Ruthenium	Ru	44	58	Curium	Cm	96	146 (150 day)
Rhodium	Rh	45	58	Berkelium	Bk	97	150 (1000 yr)
Palladium	Pd	46	60	Californium	Cf	98	153 (800 yr)
Silver	Ag	47	60	Einsteinium	Es	99	155 (480 days)
Cadmium	Cd	48	66	Fermium	Fm	100	153 (23 hr)
Indium	In	49	66	Mendelevium	Md	101	155 (1.5 hr)
Tin	Sn	50	70	Nobelium	No	102	152 (3 sec)
Antimony	Sb	51	70	Lawrencium	Lw	103	154 (8 sec)
Tellurium	Te	52	78				×/

#### Table 1

The most commonly found (most abundant in nature) isotope of each element is listed. In cases where an element has no stable isotopes, the isotope with the longest lifetime is listed.

# Isotopes

The atomic nucleus contains not only the protons we have been discussing, but also neutrons. The nuclear force, unlike the electric force, is the same between protons and neutrons, and ignores electrons. The nuclear force between nucleons (protons or neutrons), is attractive if the nucleons are close but not too close, and repulsive if you try to shove nucleons into each other. As a result, due to the nuclear force, protons and neutrons in a nucleus stick to each other forming a ball of nuclear matter as indicated in Figure (1). The nuclear force is strong enough to hold the nucleus together despite the electrical repulsion between the protons. In the nucleus of a given element, there is no precisely fixed number of neutrons. The hydrogen nucleus which has one proton, can have zero, one, or two neutrons. Helium

Figure 1a

Picture the nucleus as a spherical ball of protons and neutrons.





**Figure 1b** Model of the uranium nucleus constructed from styrofoam balls. The dark balls represent protons.

nuclei, which have 2 protons usually also have 2 neutrons, but can be found with only one neutron. Generally the light elements have roughly equal numbers of protons and neutrons while the heavy elements like uranium have a considerable excess of neutrons.

Atoms of the same element with different numbers of neutrons in the nucleus are called different **isotopes** of the element. We distinguish different isotopes of an element by appending to the name of the element a number equal to the total number of protons and neutrons in the nucleus. The hydrogen atom with just a single proton for a nucleus is hydrogen-1. If there is one neutron in addition to the proton, the atom is called hydrogen-2, and with 2 neutrons and one proton, we have hydrogen-3. These isotopes and the isotopes of helium are indicated schematically in Figure (2).

The naturally occurring isotope of uranium, an unstable element, but one with such a long half life (4.5 billion years) that it has survived since the formation of the earth, is uranium-238 or U-238. Since uranium has 92 protons, the number of neutrons in U– 238 must be 238 - 92 = 144. Another uranium isotope, U–235 with 3 fewer neutrons, is a more highly radioactive material from which atomic bombs can be constructed. (U–238 is so long–lived, so stable that it is quite safe to handle. There was some discussion of using U–238 for the keels of America's



**Figure 2** *Isotopes of hydrogen and helium.* 

cup yachts because of its very high density, but this was disallowed as being too high tech.) A list of the most common or longest lived isotopes for each element is given in Table 1.

For historical reasons, the isotopes of hydrogen are given special names. Hydrogen-2, with one proton and one neutron, is known as *deuterium*. Just over one in ten thousand hydrogen atoms in naturally occurring hydrogen are the deuterium isotope. Water molecules (H<sub>2</sub>0), in which one of the hydrogen atoms is the deuterium isotope, are called *heavy water*. Heavy water played an important role in the unsuccessful German effort to build a nuclear bomb during World War II. Hydrogen–3, with one proton and 2 neutrons, is called *tritium*. Tritium is unstable with a half life of 12.5 years. Along with deuterium, tritium plays an important role in mankind's attempt to build a nuclear fusion reactor.

Why are some isotopes stable while others are not? Why are there roughly equal numbers of neutrons as protons in the small stable isotopes and an excess of neutrons in the large ones? Why are some elements more abundant than others—for example, why does the earth have an iron core? These are questions whose answers depend upon an interplay inside the nucleus between the nuclear, the electric, and the weak interaction. We reserve a discussion of these questions for the next chapter where we discuss the nuclear and weak interactions in more detail.

# THE ELECTRIC FORCE LAW

Since electrons, protons, and neutrons make up almost everything we see around us (except for photons or light itself), a description of the electric force between these three particles provides a fairly complete picture of the electric interaction, insofar as it affects our lives. For electrons, protons, and neutrons at rest, this interaction is completely summarized in Figure (3).

As we see, protons repel each other, electrons repel each other, and a proton and an electron attract each other. There is no electric force on a neutron. The strength of the electric force between these particles drops off as  $1/r^2$  and has a magnitude shown in Equation 1. We know, to extremely high precision, that the attractive force between an electron and proton has the same strength as the repulsive force between two protons or two electrons, when the particles have the same separation r.

It is surprising how complete a summary of the electric interaction Figure (3) represents. We have only shown the forces between the particles at rest. But if you combine these results with the special theory of relativity, you can deduce the existence of magnetism and derive the formulas for magnetic forces. We will do this in Chapter 28.



# **Figure 3** *The electric interaction between protons and electrons at rest. There is no electric force on a neutron.*

#### Strength of the Electric Interaction

If two electrons or two protons are separated by a distance of one centimeter, then according to Equation 1 there is a repulsive force between them—whose strength is  $2.3 \times 10^{-19}$  dynes. Since a dyne is the weight of one milligram of mass,  $2.3 \times 10^{-19}$  dynes is a very small force. But, of course, protons and electrons are very small particles.

To get a better idea of how strong the electric force is, let us compare it with the gravitational force. If we have 2 protons any distance r apart, then the ratio of the electric repulsion to the gravitational attraction is

$$\frac{\text{electric repulsion}}{\text{gravitational attraction}} = \frac{\left|\vec{F}_{e}\right|}{\left|\vec{F}_{g}\right|}$$

$$= \frac{2.3 * 10^{-19}/\text{r}^{2}}{\text{Gm}_{p}\text{m}_{p}/\text{r}^{2}}$$
(2)

Since both electricity and gravity are  $1/r^2$  forces, the  $r^2$  cancel out in Equation 2 and we are left with

$$\frac{\left|\vec{F_{e}}\right|}{\left|\vec{F_{g}}\right|} = \frac{2.3 \times 10^{-19}}{Gm_{p}^{2}}$$
$$= \frac{2.3 \times 10^{-19}}{6.67 \times 10^{-8} \times \left(1.67 \times 10^{-24}\right)^{2}}$$

(3)

The electrical force is some  $10^{36}$  times stronger than gravity. This is true no matter how far apart the protons are. The only reason that electric forces do not completely swamp gravitational forces is that there are both attractive and repulsive electric forces which tend to cancel on a large scale, when many electrons and protons are involved.

# **ELECTRIC CHARGE**

From a historical perspective, the electric interaction was carefully studied and the electric force law well known long before the discovery of electrons and nuclei, even before there was much evidence for the existence of atoms. The simple summary of the electric force law given by Equation 1 could only be written after the 1930s, when we finally began to understand what was going on inside an atom. Prior to that, the electric force law was expressed in terms of electric charge, a concept invented by Benjamin Franklin. What we want to do in this section is to show how the concept of electric charge evolves from the forces pictured in Figure (3), and why electric charge is such a useful concept.

To convert Equation 1 into the more standard form of the electric force law, we will begin by writing the numerical constant  $2.3 \times 10^{-19}$  dyne cm<sup>2</sup>, or  $2.3 \times 10^{-20}$  newton meter<sup>2</sup>, in the form Ke<sup>2</sup> to give

$$\left| \vec{F}_{e} \right|_{electron} = \frac{Ke^2}{r^2} \quad \begin{array}{c} electric force \\ between two \\ electrons \end{array}$$
(4)

where e is called the *charge on an electron* and k is a numerical constant whose size depends on the system of units we are using.

The form of Equation 4 is chosen to make the electric force law look like the gravitational force law. To see this explicitly, compare the formulas for the magnitude of the electric and the gravitational forces between two electrons

$$\left| \vec{F}_{\text{gravitational}} \right| = \frac{Gm_{e}m_{e}}{r^{2}}$$
$$\left| \vec{F}_{\text{electric}} \right| = \frac{Kee}{r^{2}}$$
(5)

In words, we said that the gravitational force between two electrons was proportional to the product of the masses  $m_e$ , and inversely proportional to the square of the separation  $1/r^2$ . Now we say that the electric force is proportional to the product of the charges (e), and inversely proportional to the square of the distance  $1/r^2$ . By introducing the constant (e) as the **charge on the electron**, we have electric charge playing nearly the same role for the electric force law as mass does for the gravitational force law.

To get the numerical value for the charge (e) on an electron, we note that in the CGS system of units it is traditional to set the proportionality constant (K) equal to one, giving

$$\left|\vec{F}_{e}\right|\left(CGS\right) = \frac{e^{2}}{r^{2}} \qquad \begin{array}{c} electric \ force \\ law \ in \ CGS \\ units \ (K=1) \end{array}$$
(6)

Comparing Equations 1 and 6 we get

$$e^{2} = 2.3 \times 10^{-19} \text{ dynes } \text{cm}^{2}$$
$$e = 4.8 \times 10^{-10} \sqrt{\text{ dynes } \text{cm}^{2}} \qquad \begin{array}{c} charge \text{ on} \\ electron \text{ in} \\ CGS \text{ units} \end{array}$$

In the MKS system, the proportionality constant K is not 1, therefore Equation 6 does not apply to that system. Calculations involving electric forces on an atomic scale are simpler in the CGS system because of the choice K = 1. The MKS system turns out,

however, to be far more convenient when working with practical or engineering applications of electrical theory. As a result we will use the MKS system throughout the chapters on electric fields and their application, and restrict our use of the CGS system to discussions of atomic phenomena.

You will notice that the dimensions of e, displayed in Equation 4 are fairly messy. To avoid writing  $\sqrt{dynes}$  cm<sup>2</sup> all the time, this set of units is given the name *esu* which stands for electric charge as measured in the <u>*electrostatic system of units*</u>. Thus we can rewrite Equation 7 as

$$e = 4.8 \times 10^{-10} esu$$
 (7a)

as the formula for the amount of charge on an electron. (In the MKS system, electric charge is measured in *coulombs* rather than esu. The difference between a coulomb and an esu arises not only from the different set of units (newtons vs dynes) but also from the different choice of K in the MKS system. Any further discussion of the MKS system will be reserved for later chapters.

#### **Exercise 1**

(7)

What would be the value of the electric force constant K in a system of units where distance was measured in centimeters and the charge e on the electron was set equal to 1?

#### **Positive and Negative Charge**

It was Ben Franklin who introduced the concept of two kinds of electric charge. Franklin noticed that you get opposite electrical effects when you rub a glass rod with silk, or rub a rubber rod with cat fur. He decided to call the charge left on the glass rod positive charge, and the charge left on the rubber rod negative charge. What we will see in this section is how this choice of positive and negative charge leads to the electron having a negative charge – e, and a proton a positive charge + e.

charge on an electron = -e (8a)

charge on a proton = + e (8b)

Despite the fact that the electron's charge turns out to be negative, e is still called the *charge on an electron*.

The basis for saying we have two kinds of charge is the fact that with the electric interaction we have both attractive and repulsive forces. With the choice that electrons are negative and protons are positive, then the rules shown in Figure 3 can be summarized as follows: *like charges (2 protons or 2 electrons) repel, opposite charges (proton and electron) attract*. We can explain the lack of any force on the neutron by saying that the neutron has no charge that it is *neutral*.

Choosing one charge as positive and one as negative automatically gives us a reversal in the direction of the force when we switch from like to opposite charges

$$\begin{aligned} \left| \vec{F}_{e} \right|_{\text{proton proton}} &= \frac{(+e)(+e)}{r^{2}} = \frac{e^{2}}{r^{2}} \\ \left| \vec{F}_{e} \right|_{e\text{lectron electron}} &= \frac{(-e)(-e)}{r^{2}} = \frac{e^{2}}{r^{2}} \\ \left| \vec{F}_{e} \right|_{e\text{lectron proton}} &= \frac{(-e)(+e)}{r^{2}} = \frac{-e^{2}}{r^{2}} \end{aligned}$$
(9)

#### **Addition of Charge**

The concept of charge is particularly useful when we have to deal with complex structures involving many particles. To see why, let us start with the simplest electrical structure, the hydrogen atom, and gradually add more particles. We will quickly see that the electric force law, in the form  $|\vec{F}_e| = e^2/r^2$  becomes difficult to use.

In Figure (4) we have a model of the hydrogen atom consisting of a proton at the center and an electron orbiting about it. The proton sits nearly at rest at the center because it is 1836 times as massive as the electron, much as our sun sits at the center of our solar system because it is so much more massive than the planets.

The proton and electron attract each other with a force of magnitude (in CGS units) of  $|\vec{F}_e| = e^2/r^2$ . Since this is similar in form to the gravitational force between the earth and the moon, we expect the electron to travel in elliptical orbits around the proton obeying Kepler's laws. This would be exactly true if Newton's laws worked on the small scale of the hydrogen atom as they do on the larger scale of the earth-moon system.



The hydrogen atom consists of a proton at the center with an electron moving about it. The particles are held together by the attractive electric force between them.

Hydrogen atoms are approximately  $10^{-8}$  cm in diameter. Assume that the electron in the hydrogen atom in Figure (4) is traveling in a circular orbit about the proton. Use Newton's law  $\vec{F} = m\vec{a}$  and your knowledge about the acceleration  $\vec{a}$  of a particle moving in a circular orbit to predict the speed, in cm/sec, of the electron in its orbit. How does the electron's speed compare with the speed of light? (It had better be less.)

An analysis of the hydrogen atom is easy and straight forward using the force law  $|\vec{F}_e| = e^2/r^2$ . But the analysis gets more difficult as the complexity of the problem increases. Suppose, for example, we have two hydrogen atoms separated by a distance r. Let r be quite a bit larger than the diameter of a hydrogen atom, as shown in Figure (5a).

Even though r is much larger than the size of the individual hydrogen atoms, there are still electric forces between the protons and electrons in the two atoms. The two protons repel each other with a force  $\vec{F}_{pp}$ , the electrons repel each other with a force  $\vec{F}_{ee}$ , the proton in the left atom attracts the electron in the right atom with a force  $\vec{F}_{ep}$ . Sa you can see, eight separate forces are involved, as shown in Figure (5b).



(a) Two hydrogen atoms separated by a distance r



(b) Forces between the particles in the two atoms  $|\vec{F}_{pp}| = |\vec{F}_{ee}| = |\vec{F}_{pe}| = e^2/r^2$ 

#### Figure 5

When we have two hydrogen atoms fairly far apart, then there is essentially no net force between the atoms. The reason is that the repulsive forces between like particles in the two atoms are cancelled by the attractive force between oppositely charged particles. If the separation r of the atoms is large compared to the diameter of each hydrogen atom, then all these eight forces have essentially the same magnitude  $e^2/r^2$ . Since half are attractive and half are repulsive, they cancel and we are left with no net force between the atoms.

#### Exercise 3

A complete carbon atom has a nucleus with 6 protons, surrounded by 6 orbiting electrons. If you have two complete separate carbon atoms, how many forces are there between the particles in the two different atoms?

With just two simple hydrogen atoms we have to deal with 8 forces in order to calculate the total force between the atoms. If we have to deal with something as complex as calculating the force between two carbon atoms, we have, as you found by doing Exercise 3, to deal with 72 forces. Yet the answer is still zero net force. There must be an easier way to get this simple result.

The easier way is to use the concept of *net charge* Q which is the *sum of the charges in the object*. A hydrogen atom has a net charge

$$Q_{hydrogen} = (+e)_{proton} + (-e)_{electron}$$
$$= 0$$
(10)

The *net force* between two objects with net charges  $Q_1$  and  $Q_2$  is simply

$$\left| \vec{F}_{net} \right| = \frac{KQ_1Q_2}{r^2} \quad \begin{array}{c} Coulomb's \ law, \ where}{K = 1 \ for \ CGS \ units} \quad (11)$$

Equation 1, which looks very much like Newton's law of gravity, except that charge replaces mass, is known as *Coulomb's law*. The proportionality constant K is 1 for CGS units.

Applying Coulomb's law to the force between two complete carbon atoms, we see immediately that the complete atoms have zero net charge, and therefore by Coulomb's law there is no net force between them. The cancellation of individual forces seen in Figure (4) is accounted for by the cancellation of charge in Coulomb's law.

To see that the addition of charge and Coulomb's law work in situations where charge does not cancel, suppose we had two helium nuclei separated by a distance r. (A bare helium nucleus, which is a helium atom missing both its electrons, would be called a *doubly ionized* helium atom.) In Figure (6) we have sketched the forces between the two protons in each nucleus. Both protons in nucleus number 1 are repelled by both protons in nucleus number 2, giving rise to a net repulsive force four times as strong as the force between individual protons, or a force of magnitude  $4e^2/r^2$ .

Applying Coulomb's law to these two nuclei, we see that the charge on each nucleus is 2e, giving for the charges  $Q_1$  and  $Q_2$ 

 $Q_1 = 2e = total charge on nucleus #1$ 

 $Q_2 = 2e = total charge on nucleus #2$ 

Thus Coulomb's law gives (with K = 1 for CGS units)

$$\left|\vec{F}_{e}\right| = \frac{Q_{1}Q_{2}}{r^{2}} = \frac{(2e)(2e)}{r^{2}} = \frac{4e^{2}}{r^{2}}$$
 (12)

which is in agreement with Figure (6).



Figure 6

We see that the repulsive force between 2 helium nuclei is 4 times as great as the repulsion  $e^2/r^2$ between 2 protons. Using Coulomb's law  $|\vec{F}| = Q_1 Q_2/r^2$  with  $Q_1 = Q_2 = 2$  e for the helium nuclei gives the same result.

#### Exercise 4

This exercise is designed to give you a more intuitive feeling for the enormous magnitude of the electric force, and how complete the cancellation between attractive and repulsive forces is in ordinary matter.

Imagine that you could strip all the electrons from two garden peas, leaving behind two small balls of pure positive charge. Assume that there is about one mole  $(6 \times 10^{23})$  protons in each ball.

(a) What is the total charge Q on each of these two balls of positive charge? Give the answer in esu.

(b) The two positively charged peas are placed one meter (100 cm) apart as shown. Use Coulomb's law



to calculate the magnitude of the repulsive force between them. Give your answer in dynes and in metric tons. (1 metric ton =  $10^9$  dynes  $\approx$  1 english ton.)

If you worked Exercise 4 correctly, you found that two garden peas, stripped of all electrons and placed one meter apart, would repel each other with a force of nearly  $10^{16}$  tons! Yet when you actually place two garden peas a meter apart, or only a centimeter apart, there is no observable force between them. The  $10^{16}$  ton repulsive forces are so precisely cancelled by  $10^{16}$  ton attractive forces between electrons and protons that not even a dyne force remains.

#### Exercise 5

What would be the repulsive force between the peas if only one in a billion (one in 10<sup>9</sup>) electrons were removed from each pea, and the peas were placed one meter apart?
#### **CONSERVATION OF CHARGE**

Up to this point, we have used the concept of electric charge to simplify the calculation of the electric force between two objects containing many electrons and protons. But the fact that electrons and protons have precisely opposite charges suggests that in nature electric charge has a deeper significance. That deeper significance is the conservation of electric charge. Like the conservation of energy, linear momentum, and angular momentum, the *conservation of electric charge* appears to be a basic law with no known exceptions.

When we look beyond the familiar electrons and protons, into the world of subnuclear particles, we find a bewildering array of hundreds of different kinds of particles. In the chaos of such an array of particles, two features stand out. Almost all of the particles are unstable, and when the unstable particles decay, electric charge is conserved. In looking at the particle decays, it becomes clear that there really is something we call electric charge that is passed from one particle to another, and not lost when a particle decays.

We will illustrate this with a few examples. We have already discussed several unstable particles, the muon introduced in the muon lifetime experiment, the  $\pi$ mesons, created for cancer research, and the neutron which, by itself outside a nucleus, has a half life of nine minutes. The muon decays into an electron and a neutrino, and the neutron decays into a proton, electron and an antineutrino (the antiparticle of the neutrino). There are three separate  $\pi$  mesons. The negative charged one decays into an electron and an antineutrino, the positive one into a positron (antielectron) and neutrino, and the neutral one into two photons.

We can shorten our description of these decays by introducing shorthand notation for the particles and their properties. We will use the Greek letter  $\mu$  (mu) for the muon,  $\pi$  for the  $\pi$ mesons,  $\nu$  (nu) for the neutrino and  $\gamma$  (gamma) for protons. We designate the charge of the particle by the superscript + for a positive charge, – for a negative, and 0 for uncharged. Thus the three  $\pi$ mesons are designated  $\pi^+$ ,  $\pi^0$ , and  $\pi^-$  for the positive, neutral and negative ones respectively. In later discussions, it will be useful to know whether we are dealing with a particle or an antiparticle. We denote antiparticles by putting a bar over the symbol, thus v represents a neutrino, and  $\nabla$  an antineutrino. Since a particle and an antiparticle can annihilate each other, a particle and an antiparticle must have opposite electric charges if they carry charge at all, so that charge will not be lost in the annihilation. As a result the antiparticle of the electron e<sup>-</sup> is the positively charged positron which we designate  $\bar{e}^+$ .

Using these conventions, we have the following notation for the particles under discussion (photons and neutrinos are uncharged):

Notation	Particle		Particle
p+	proton	$ abla^0$	antineutrino
$n^0$	neutron	μ-	muon
e-	electron	$\pi^+$	pi plus
ē+	positron	$\pi^0$	pi naught
$\gamma^0$	photon	$\pi^{-}$	pi minus
$v^0$	neutrino		

The particle decays we just described can now be written as the following reactions.

$\mu^-  ightarrow e^- + \nabla^0$	muon decay	(a)
$n \rightarrow p^+ + e^- + \overline{v}^0$	neutron decay	(b)
$\pi^- \rightarrow \bar{e}^+ + \nu^0$	pi minus decay	(c)
$\pi^0 \rightarrow \gamma^0 + \gamma^0$	pi naught decay	(d)
$\pi^- \rightarrow e^- + \nabla^0$	pi minus decay	(e) (14)

Note that in all of these decays, the particles change but the charge does not. If we start with a negative charge, like the negative muon, we end up with a negative particle, the electron. If we start with a neutral particle like the  $\pi^0$ , we end up with no net charge, in this case two photons.

Among the hundreds of elementary particle decays that have been studied, no one has found an example where the total charge changed during the process. It is rather impressive that the concept of positive and negative charge, introduced by Ben Franklin to explain experiments involving rubber rods and cat fur, would gain even deeper significance at the subnuclear level.

#### **Stability of Matter**

The conservation of electric charge may be related to the stability of matter. The decay of elementary particles is not an exceptional occurrence, it is the general rule. Of the hundreds of particles that have been observed, only four are stable, the proton, the electron, the photon and the neutrino. (Neutrons are also stable if buried inside a nucleus, for reasons we will discuss in the next chapter.) All the other particles eventually, and often very quickly, decay into these four.

The question we should ask is not why particles decay, but instead why these four particles do not. We know the answer in the case of two of them. Photons, and perhaps, neutrinos, have zero rest mass. As a result they travel at the speed of light, and time does not pass for them. If a photon had a half life, that half life would become infinite due to time dilation.

Why is the electron stable? It appears that the stability of the electron is due to the conservation of energy and electric charge. The electron is the least massive charged particle. There is nothing for it to decay into and still conserve charge and energy.

That leaves the proton. Why is it stable? We do not know for sure. There are a couple of possibilities which are currently under study. One is that perhaps the proton has some property beyond electric charge that is conserved, and that the proton is the least massive particle with this property. This was the firm belief back in the 1950s.

In the 1960s, with the discovery of quarks and the combining of the electric and weak interaction theories, it was no longer obvious that the proton was stable. Several theories were proposed, theories that attempted to unify the electric, weak, and nuclear force. These so-called *Grand Unified Theories* or *GUT* for short, predicted that protons should eventually decay, with a half life of about  $10^{31}$  years. Since the universe is only  $10^{10}$  years old, that is an incredibly long time.

It is not impossible to measure a half life of  $10^{31}$  years. You do not have to wait that long. Instead you look at  $10^{31}$  or  $10^{32}$  particles, and see if a few decay in one year. Since a mole of particles is  $6 \times 10^{23}$  particles, you need about a billion moles of protons for such an experiment. A mole of protons (hydrogen) weighs one gram, a billion moles is a million kilograms or a thousand metric tons. You get that much mass in a cube of water 10 meters on a side, or in a large swimming pool. For this reason, experiments designed to detect the decay of the proton had to be able to distinguish a few proton decays per year in a swimming pool sized container of water.

So far none of these detectors has yet succeeded in detecting a proton decay (but they did detect the neutrinos from the 1987 supernova explosion). We now know that the proton half-life is in excess of  $10^{32}$  years, and as a result the Grand Unified Theories are in trouble. We still do not know whether the proton is stable, or just very long lived.

#### **Quantization of Electric Charge**

Every elementary particle that has been detected individually by particle detectors has an electric charge that is an integer multiple of the charge on the electron. Almost all of the particles have a charge + e or - e, but since the 1960s, a few particles with charge 2e have been observed. Until the early 1960s it was firmly believed that this *quantization* of charge in units of + e or - e was a basic property of electric charge.

In 1961, Murray Gell-Mann, who for many years had been trying to understand the bewildering array of elementary particles, discovered a symmetry in the masses of many of the particles. This symmetry, based on the rather abstract mathematical group called SU2, predicted that particles with certain properties, could be grouped into categories of 8 or 10 particles. This grouping was not unlike Mendeleev's earlier grouping of the elements in the periodic table.

When the periodic table was first constructed, there were gaps that indicated missing, as yet undiscovered elements. In Gell-Mann's SU3 symmetry there were also gaps, indicating missing or as yet undetected elementary particles. In one particular case, Gell-Mann accurately predicted the existence and the properties of a particle that was later discovered and named the  $\omega^{-}$ (omega minus). The discovery of the  $\omega^{-}$ verified the importance of Gell-Mann's symmetry scheme.

In 1964 Gell-Mann, and independently George Zweig also from Caltech, found an exceedingly simple model that would explain the SU3 symmetry. They found that if there existed three different kinds of particles which Gell-Mann called *quarks*, then you could make up *all* the known heavy elementary particles out of these three quarks, and the particles you make up would have just the right SU3 symmetry properties. It was an enormous simplification to explain hundreds of "elementary particles" in terms of 3 kinds of quarks.

Our discussion of quarks will be mainly reserved for the next chapter. But there is one property of quarks that fits into our current discussion of electric charge. The charge on a quark or anti quark can be + 1/3 e, -1/3 e, +2/3 e or -2/3 e. The charge e on the electron turns out not to be the fundamental unit of charge.

An even stranger property of quarks is the fact that they exist only inside elementary particles. For example, a proton or neutron is made up of three quarks and a  $\pi$  meson of two. All particles made from quarks have just the right number of quarks, in just the right combination, so that the total charge of the particle is an integer multiple of the electron charge e. Although the quarks themselves have a fractional charge  $\pm 1/3$  e or  $\pm 2/3$  e, they are always found in combinations that have an integer net charge.

You might ask, why not just tear a proton apart and look at the individual quarks? Then you would see a particle with a fractional charge. It now appears that, due to an unusual property of the so-called *color force* between quarks, you cannot simply pull quarks out of a proton. The reason is that the color force, unlike gravity and electricity, becomes stronger, not weaker, as the separation of the particles increases. We will see later how this bizarre feature of the color force makes it impossible to extract an individual quark from a proton.

#### **MOLECULAR FORCES**

A naïve application of Coulomb's law would say that complete atoms do not interact. A complete atom has as many electrons outside as protons in the nucleus, and thus zero total charge. Thus by Coulomb's law, which says that the electric force between two objects is proportional to the product of the charges on them, one would predict that there is no electric force between two complete atoms. Tell that to the two hydrogen atoms that bind to form hydrogen molecules, oxygen atoms that bind to form to  $O_2$  molecules we breathe, or the hydrogen and oxygen atoms that combine to form the water molecules we drink. These are all complete atoms that have combined together through electric forces to form molecular structures.

The reason that neutral atoms attract electrically to form molecules is the fact that the negative charge in an atom is contained in the electrons which are moving about the nucleus, and their motion can be affected by the presence of other atoms.

When trying to understand molecular forces, the planetary picture of an atom, with electrons in orbits like the planets moving around the sun, is not a particularly useful or accurate model. A more useful picture, which has its origin both in quantum mechanics and Newtonian mechanics is to picture the electrons as forming a cloud of negative charge surrounding the nucleus. You can imagine the electrons as moving around so fast that, as far as neighboring atoms are concerned, the electrons in an atom simply fill up a region around the nucleus with negative electric charge. When doing accurate calculations with quantum mechanics, one finds that the electron clouds have definite shapes, shapes which the chemists call *orbitals*.

One does not need quantum mechanics in order to get a rough understanding of the origin of molecular forces. Simple arguments about the behavior of the electron clouds gives a fairly good picture of what the chemists call *covalent bonding*. We will illustrate this with a discussion of the hydrogen molecule.

#### **Hydrogen Molecule**

To construct a hydrogen molecule, imagine that we start with a single proton and a complete hydrogen atom as indicated in Figure (7a). Here we are representing the hydrogen atom by a proton with the electron moving around to more or less fill a spherical region around the proton. In this case the external proton is attracted to the sphere of negative charge by a force that is as strong as the repulsion from the hydrogen nucleus. As a result there is very little force between the external proton and the neutral atom. Here, Coulomb's law works.

Now bring the external proton closer to the hydrogen atom, as shown in Figure (7b). Picture the hydrogen nucleus as fixed, nailed down, and look at what happens to the hydrogen electron cloud. The electron is now beginning to feel the attraction of the external proton as well as its own proton. The result is that the electron cloud is distorted, sucked over a bit toward the external proton. Now the center of the electron cloud is a bit closer to the external proton than the hydrogen nucleus, and the attractive force between the electron cloud and the proton is slightly stronger than the repulsive force between the protons. The external proton now feels a net attractive force to the neutral hydrogen atom because of the distortion of the electron cloud. A naïve application of Coulomb's law ignores the distortion of the electron cloud and therefore fails to predict this attractive force.

Since the external proton in Figure (7b) is attracted to the hydrogen atom, if we let go of the proton, it will be sucked into the hydrogen atom. Soon the electron will start orbiting about both protons, and the external proton will be sucked in until the repulsion between the protons just balances the electrical attraction. Since the two protons are identical, the electron will have no reason to prefer one proton over the other, and will form a symmetric electron cloud about both protons as shown in Figure (7c). The result is a complete and stable object called a *hydrogen molecule ion*.



(b) The external proton is brought closer distorting the electron cloud





(c) Electron orbits both protons

Hydrogen molecule ion

#### Figure 7a

Formation of a hydrogen molecule ion. To visualize how a hydrogen molecule ion can be formed, imagine that you bring a proton up to a neutral hydrogen.

#### Figure 7b

When the proton gets close, it distorts the hydrogen electron cloud. Since the distorted cloud is closer to the external proton, there is a net attractive force between the proton and the distorted hydrogen atom.

#### Figure 7c

If the protons get too close, they repel each other. As a result there must be some separation where there is neither attraction or repulsion. This equilibrium separation for the protons in a hydrogen molecule ion is 1.07 Angstroms. (1 Angstrom =  $10^{-8}$  cm.) The final step in forming a hydrogen molecule is to note that the hydrogen molecule ion of Figure (7c) has a net charge +e, and therefore will attract another electron. If we drop in another electron, the two electrons form a new symmetric cloud about both protons and we end up with a stable  $H_2$  molecule, shown in Figure (8).

Although the discussion related to Figures (7, 8) is qualitative in nature, it is sufficient to give a good picture of the difference in character between atomic

and molecular forces. Atomic forces, the pure  $e^2/r^2$ Coulomb force that binds electrons to the nucleus, is very strong and fairly simple to understand. Molecular forces, which are also electrical in origin but which depend on subtle distortion of the electron clouds, are weaker and more complex. Molecular forces are so subtle that you can make very complex objects from them, for example, objects that can read and understand this page. The sciences of chemistry and biology are devoted primarily to understanding this complexity.



#### Figure 8

The hydrogen molecule ion of Figure (7c) has a net positive charge +e, and therefore can attract and hold one more electron. In that case both electrons orbit both protons and we have a complete hydrogen molecule. The equilibrium separation expands to 1.48 Angstroms.

#### Molecular Forces—A More Quantitative Look

It is commonly believed that quantum mechanics, which can be used to predict the detailed shape of electron clouds, is needed for any quantitative understanding of molecular forces. This is only partly true. We can get a fair understanding of molecular forces from Newtonian mechanics, as was demonstrated by the student Bob Piela in a project for an introductory physics course. This section will closely follow the approach presented in Piela's project.

In this section we will discuss only the simplest of all molecules, the hydrogen molecule ion consisting of two protons and one electron and depicted in Figure (7c). We will use Newtonian mechanics to get a better picture of how the electron holds the molecule together, and to see why the lower, the more negative the energy of the electron, the more tightly the protons are bound together.

If you do a straightforward Newtonian mechanics calculation of the hydrogen molecule ion, letting all three particles move under the influence of the Coulomb forces between them, the system eventually flies apart. As a number of student projects using computer calculations have shown, eventually the electron gets captured by one of the protons and the other proton gets kicked out of the system. With Newtonian mechanics we cannot explain the stability of the hydrogen molecule ion, quantum mechanics is required for that.

Piela avoided the stability problem by assuming that the two protons were fixed at their experimentally known separation of  $1.07 \times 10^{-8}$  cm (1.07 angstroms) as shown in Figure (9a), and let the computer calculate the orbit of the electron about the two fixed protons, as seen in Figure (9b). By letting the calculation run for a long time and plotting the position of the electron at equal intervals of time, as in a strobe photograph of the electron's motion, one obtains the dot pattern seen in Figure (9c). This dot pattern can be thought of as the classical electron cloud pattern for the electron.



a) Electric force acting on the electron.



b) Line drawing plot of the orbit of the electron about the two protons



c) Dots showing the position of the electron at equal time intervals (effectively a strobe photograph).

#### Figure 9

Orbit of an electron about two fixed protons.

#### **The Bonding Region**

Of the dots we have drawn in Figure (9c) some are more effective than others in holding the molecule together. When the electron is between the protons, it pulls in on both protons providing a net bonding force. But when the electron is outside to the left or right, it tends to pull the protons apart. We can call the region where the electron gives rise to a net attractive force the bonding region, while the rest of space, where the electron tends to pull the protons apart, can be called the anti-bonding region.

To see how we can distinguish the bonding from the anti-bonding region, consider Figure (10) where we show the forces the electron exerts on the protons for several positions of the electron. In (10a), the electron is between but above the protons, giving rise to the forces  $\vec{F}_1$  and  $\vec{F}_2$  shown. We also show the x components which are  $\vec{F}_{1x}$  and  $\vec{F}_{2x}$ . These components are of more interest to us than  $\vec{F}_1$  and  $\vec{F}_2$  because the electron, while in orbit, will spend an

equal time above and below the protons. Thus on the average the y components cancel, and the net effect of the electron is described by the x components  $\vec{F}_{1x}$  and  $\vec{F}_{2x}$  alone. We can see from Figure (10a) that  $\vec{F}_{1x}$  and  $\vec{F}_{2x}$  are pulling the protons together. This electron is clearly in the bonding region.

It is a little bit harder to see the anti-bonding forces. In Figure (10b), we show the electron first to the left of the protons, then to the right. Again we concentrate on the x components  $\vec{F}_{1x}$  and  $\vec{F}_{2x}$  because the y components will, on the average, cancel. However, for orbits like that shown in Figure (9b), the electron spends the same amount of time on the left as the right, and thus we should average  $\vec{F}_{1x}$  and  $\vec{F}_{2x}$  for these two cases. When we do this, we see that the average  $\vec{F}_{1x}$  points left, the average  $\vec{F}_{2x}$  points right, and these two average forces are pulling the protons apart.

Average x component of force



#### Figure 10

average out since the electron spends as much time above as below the protons.)

Bonding and anti bonding regions. When the electron is in the bonding region, the electric force exerted by the electron on the protons pulls the protons together. When in the anti-bonding region, the electric force pulls the protons apart. If you are calculating an orbit and want to test whether the electron is in the bonding or antibonding region, simply compare  $\vec{F}_{1x}$  and  $\vec{F}_{2x}$ . If the electron is to the right of the protons and  $\vec{F}_{1x}$  is bigger than  $\vec{F}_{2x}$ , the protons are being pulled apart. If you are to the left of the protons, and  $\vec{F}_{2x}$  is bigger than  $\vec{F}_{1x}$ , then again the protons are being pulled apart. Otherwise the protons are being pulled together and the electron is in the bonding region.

In Figure (11), we replotted the electron dot pattern of Figure (9c), but before plotting each point, checked whether the electron was in the bonding or antibonding region. If it were in the bonding region, we plotted a dot, and if it were in the anti-bonding region we drew a cross. After the program ran for a while, it became very clear where to draw the lines separating the bonding from the anti-bonding regions.



a) Electron cloud for a -10eV electron. Points in the bonding region are plotted as dots, outside in the anti-bonding region as crosses.



**Figure 12** *Determining the bonding region from the computer plot.* 

#### **Electron Binding Energy**

One of the features of Figure (12) is that there are many dots out in the anti-bonding region. It looks as if there are more dots out there pulling the protons apart than inside holding them together. In this case is the electron actually helping to hold the molecule together?

One way to tell whether or not a system will stay together or fall apart is to look at the total energy of the system. If it costs energy to pull a system apart, it will stay together. But if energy is released when a system comes apart, it will fall apart.

In our earlier discussion in Chapter 8 of the motion of a satellite around the earth, we saw that if the total energy of the satellite were negative, the satellite would be bound to the earth and could not escape. On the other hand, if the total energy of the satellite were positive, the satellite would eventually escape no matter what direction it was heading (assuming it did not crash). These predictions about total energy applied if we did not include rest energy, and assumed that the gravitational potential energy was zero when the satellite and earth were infinitely far apart. This gave us as the formula for gravitational potential energy

$$E_{tot} \begin{bmatrix} planet \\ and \\ satellite \end{bmatrix} = \frac{1}{2}m_s v^2 - \frac{Gm_s M_e}{r}$$
(8-29)

where  $m_s$  and  $M_e$  are the masses of the satellite and earth, and v the speed of the satellite.

In describing the motion of an electron in an atom or molecule, we can use the convention that the electron's electric potential energy is zero when it is infinitely far away from the protons. With this convention, the formula for the electric potential energy between an electron and a proton a distance r apart is  $-Ke^2/r$ , which is analogous to the gravitational potential energy  $-Gm_sM_e/r$ . (Simply replace  $Gm_sM_e$  by  $Ke^2$  to go from a discussion of gravitational forces in satellite motion to electric forces in atoms.) Thus the formula for the total energy of an electron in orbit about a single stationary proton should be (in analogy to Equation 8-29)

$$E_{tot} \begin{bmatrix} electron \\ and \\ proton \end{bmatrix} = \frac{1}{2} m_e v^2 - \frac{Ke^2}{r}$$
(15)

where  $m_e$  is the mass of the electron, v its speed, and r the separation between the electron and proton.

Just as in the case of satellite motion, the total energy tells you whether the electron is bound or will eventually escape. If the total energy is negative, the electron cannot escape, while if the total energy is positive, it must escape.

Another way of describing the electron's behavior is to say that if the electron's total energy is negative, it is down in some kind of a well and needs outside help, outside energy, in order to escape. The more negative the electron's total energy, the deeper it is in the well, and the more tightly bound it is. We can call the amount by which the electron's total energy is negative the *binding energy* of the electron. The binding energy is the amount of energy that must be supplied to free the electron.

#### Electron Volt as a Unit of Energy

In discussing the motion of an electron in an atom, quantities like meters and kilograms and joules are awkwardly large. There is, however, a unit of energy that is particularly convenient for discussing many applications, including the motion of electrons in atoms. This unit of energy, called the *electron volt* (abbreviated eV), is the amount of energy an electron would gain if it hopped from the negative to the positive terminal of a 1 volt battery. The numerical value is

1 electron volt (eV) = 
$$1.6 \times 10^{-19}$$
 joules  
=  $1.6 \times 10^{-12}$  ergs (16)

As an example, an electron in a cold (unexcited) hydrogen atom has a total energy of -13.6 eV. The fact that the electron's energy is negative means that the electron is bound to the proton–cannot escape.

The value -13.6 eV means that, in order to pull the electron out of the hydrogen atom, we would have to supply 13.6 eV of energy. In other words, the binding energy of the electron in a cold hydrogen atom is 13.6 eV. The number is 13.6 eV is much easier to discuss and remember than  $2.16 \times 10^{-18}$  joules.

Another unit that is convenient for discussing atoms is the *angstrom* (abbreviated  $\mathring{A}$ ) which is  $10^{-8}$  cm or  $10^{-10}$  meters.

1 Angstrom(Å) = 
$$10^{-8}$$
cm  
=  $10^{-10}$ m (17)

A hydrogen atom has a diameter of  $1 \text{ \AA}$  and all atoms are approximately the same size. Even the largest atom, Uranium, has a diameter of only a few angstroms. In the hydrogen molecule ion, the separation of the protons is  $1.07 \text{ \AA}$ .

#### Electron Energy in the Hydrogen Molecule Ion

We have seen that the strength of the binding of an electron in an atom is related to the total energy of the electron. The more negative the energy of the electron, the more tightly it is bound.

Let us now return to our discussion of the hydrogen molecule ion to see if the total energy of the electron in its orbit about the two protons is in any way related to the effectiveness which the electron binds the proton together.

When the electron is orbiting about two protons, there are two electric potential energy terms, one for each proton. Thus the formula for the electron's total energy is

$$E_{tot}\begin{bmatrix} electron\\ in H_2^+\\ molecule \end{bmatrix} = \frac{1}{2}m_e v^2 - \frac{Ke^2}{r_1} - \frac{Ke^2}{r_2}$$
(18)

which is the same as Equation 15 except for the additional potential energy term. In this equation,  $r_1$  is the distance from the electron to proton #1, and  $r_2$  to proton #2.

When we wrote computer programs for satellite motion, we found that it was much easier to work in a system of units where the earth mass, earth radius and hour were set to 1. In these units the gravitational constant G was simply 20, and we never had to work with awkwardly large numbers.

Similarly, we can simplify electron orbit calculations by choosing a set of units that are convenient for these calculations. In what we will call *atomic units*, we will set the mass  $m_e$  of the electron, the electric charge e, the angstrom, and the electron volt all to 1. When we do this, the electric force constant K has the simple value of 14.40. These choices are summarized in Table 2.

Using atomic units, the formula for the total energy of the electron in the hydrogen molecule ion (Equation 18) reduces to

$$E_{tot}(H_2^+) = \frac{v^2}{2} - \frac{14.40}{r_1} - \frac{14.40}{r_2}$$
(18a)

since  $m_e = e = 1$ . The real advantage of this formula is that it directly gives the electron's total energy in electron volts, no conversion is required. Because energy is conserved, because the electron's total energy does not change as the electron goes around in its orbit, we can name the orbit by  $E_{tot}$ . For example, the orbit shown back in Figures (9) and (11), had a total energy of -10 eV. We can say that this was a "-10 eV orbit".

When Bob Piela did his project on the hydrogen molecule ion, his main contribution was to show how the energy of the electron in orbit was related to the bonding force exerted by the electron on the protons. Piela's results are easily seen in Figure (13). In (13a), we show the -10 eV orbit superimposed upon a sketch of the bonding region. In (13b), the same orbit is shown as a strobe photograph. As we mentioned earlier, there appear to be a lot more dots outside the bonding region than inside, and it does not look like a -10 eV electron does a very good job of binding the protons in the H<sup>+</sup><sub>2</sub> molecule.

In Figure (13c) we have plotted a -20 eV orbit. The striking feature is that as the electron energy is reduced, made more negative, the electron spends more time in the bonding region doing a better job of holding the molecule together. In Figure (13d) the electron energy is dropped to -30 eV and the majority of the electron cloud is now in the bonding region. It is easy to see that at -30 eV the electron does a good job of binding the protons. With Piela's diagrams it is easy to see how the electron bonds more strongly when its energy is lowered.

Symbol	Atomic Units	MKS Units
eV	1	$1.6 \times 10^{-19}$ joules
${\rm \overset{o}{A}}$	1	$10^{-10}$ m
m <sub>e</sub>	1	$9.1 \times 10^{-31} \text{kg}$
e	1	$1.6 \times 10^{-19}$ coulombs
K	14.40	$9 \times 10^9$ m/ farad
r <sub>b</sub>	.51 Å	
$I_2^+$	$1.07{ m \AA}$	
	Symbol eV Å $m_e$ e K $r_b$ $I_2^+$	Symbol         Atomic Units $eV$ 1 $\mathring{A}$ 1 $m_e$ 1 $e$ 1 $k$ 14.40 $r_b$ .51 $\mathring{A}$ $I_2^+$ 1.07 $\mathring{A}$

#### **ATOMIC UNITS**

Table 2



#### Figure 13

As we decrease the electron's total energy, the electron spends more time in the bonding region, with the result that the protons are more tightly bound. Thus we see that the lower the electron energy, the stronger the binding.

the stronger the bonding.

#### ! HYDROGEN MOLECULE ION



LET F1y = -(Ry/R)\*F1 LET F2x = -(Sx/S)\*F2 !points in -S direction LET F2y = -(Sy/S)\*F2LET Fx = F1x + F2x!Vector sum of forces LET Fy = F1y + F2y!Newton's Second law LET Ax = Fx/Me LET Ay = Fy/Me LET Vold = SQR( $Vx^*Vx + Vy^*Vy$ ) LET  $Vx = Vx + Ax^*dt$ LET  $VY = Vy + Ay^*dt$ LET Vnew = SQR(Vx\*Vx + Vy\*Vy) LET V = (Vold + Vnew)/2  $\mathsf{LET} \mathsf{T} = \mathsf{T} + \mathsf{dt}$ LET i = i + 1IF MOD(i,50) = 0 THEN PLOT Rx,Ry IF Rx > Zx THEN IF -F2x > -F1x THEN CALL CROSS END IF IF Rx < 0 THEN IF -F1x < -F2x THEN CALL CROSS END IF END IF LOOP UNTIL T > 100 ! ------ Subroutine ENERGY prints out total energy. SUB ENERGY LET Etot = Me\*V\*V/2 - K\*Qe\*Qp/R - K\*Qe\*Qp/S Add potential energy of repulsive core LET Etot = Etot + (1/2)\*Kr\*Qe\*Qp/R^2 + (1/2)\*Kr\*Qe\*Qp/S^2 PRINT T.Etot END SUB ! ----- Subroutine CROSS draws a cross at Rx,Ry. SUB CROSS PLOT LINES: Rx-.01,Ry; Rx+.01,Ry PLOT LINES: Rx,Ry-.01; Rx,Ry+.01 END SUB ! ------ Subroutine BigCROSS draws a cross at Rx,Ry. SUB BigCROSS PLOT LINES: Rx-.04, Ry; Rx+.04, Ry PLOT LINES: Rx,Ry-.04; Rx,Ry+.04 END SUB END

!Calculate force

LET F1x = -(Rx/R)\*F1 !points in -R direction

proton proton

**Figure 14** *Computer program for the hydrogen molecule ion.* 

proton

ź

proton

## Chapter 20 Nuclear Matter

In the last chapter our focus was on what one might call electronic matter—the structures that result from the interaction of the electrons in atoms. Now we look at nuclear matter, found both in the nuclei of atoms and in neutron stars. The structures we see result from an interplay of the basic forces of nature. In the atomic nucleus, the nuclear, electric, and weak interactions are involved. In neutron stars and black holes, gravity also plays a major role.

#### **NUCLEAR FORCE**

In 1912 Ernest Rutherford discovered that all the positive charge of an atom was located in a tiny dense object at the center of the atom. By the 1930s, it was known that this object was a ball of positively charged protons and electrically neutral neutrons packed closely together as illustrated in Figure (19-1) reproduced here. Protons and neutrons are each about  $1.4 \times 10^{-13}$  cm in diameter, and the size of a nucleus is essentially the size of a ball of these particles. For example, iron 56, with its 26 protons and 30 neutrons, has a diameter of about 4 proton diameters. Uranium 235 is just over 6 proton diameters across. (One can check, for example, that a bag containing 235 similar marbles is about six marble diameters across.)

That the nucleus exists means that there is some force other than electricity or gravity which holds it together. The protons are all repelling each other electrically, the neutrons are electrically neutral, and the attractive gravitational force between protons is some  $10^{-38}$  times weaker than the electric repulsive force. The force that holds the nucleus together must be attractive and even stronger than the electric repulsion. This attractive force is called the *nuclear force*.

The nuclear force treats protons and neutrons equally. In a real sense, the nuclear force cannot tell the difference between a proton and a neutron. For this reason, we can use the word nucleon to describe either a proton or neutron, and talk about the nuclear force between nucleons. Another feature of the nuclear force is that it ignores electrons. We could say that electrons have no *nuclear charge*.

The properties of the nuclear force can be deduced from the properties of the structures it creates—namely atomic nuclei. The fact that protons and neutrons maintain their size while inside a nucleus means that the nuclear force is both attractive and repulsive. Try to pull two nucleons apart and the attractive nuclear force holds them together, next to each other. But try to squeeze two nucleons into each other and you encounter a very strong repulsion, giving the nucleons essentially a solid core.

We have seen this kind of behavior before in the case of molecular forces. Molecular forces are attractive, holding atoms together to form molecules, liquids and crystals. But if you try to push atoms into each other, try to compress solid matter, the molecular force becomes repulsive. It is the repulsive part of the molecular force that makes solid matter hard to compress, and the repulsive part of the nuclear force that makes nuclear matter nearly incompressible.



**Figure 19-1a** Sketch of an atomic nucleus, showing it as a ball of protons and neutrons.



**Figure 19-1b** Styrofoam model of a Uranium nucleus. (The dark balls represent protons.)

#### **Range of the Nuclear Force**

While the attractive nuclear force must be stronger than the electric force to hold the protons together in the nucleus, it is not a long range  $1/r^2$  force like electricity and gravity. It drops off much more rapidly than  $1/r^2$ , with the result that if two protons are separated by more than a few proton diameters, the electric repulsion becomes stronger than the nuclear attraction. The separation  $R_0$  at which the electric repulsion becomes stronger than the nuclear attraction, is about 4 proton diameters. This distance  $R_0$ , which we will call the *range of the nuclear force*, can be determined by looking at the stability of atomic nuclei.

If we start with a small nucleus, and keep adding nucleons, for a while the nucleus becomes more stable if you add the right mix of protons and neutrons. By more stable, we mean more tightly bound. To be explicit, the more stable, the more tightly bound a nucleus, the more energy that is required, per nucleon, to pull the nucleus apart. This stability, this tight binding, is caused by the attractive nuclear force between nucleons.

Iron 56 is the most stable nucleus. It takes more energy per nucleon to take an Iron 56 nucleus apart than any other nucleus. If the nucleus gets bigger than Iron 56, it becomes less stable, less tightly bound. If a nucleus gets too big, bigger than a Lead 208 or Bismuth 209 nucleus, it becomes unstable and decays by itself.

The stability of Iron 56 results from the fact that an Iron 56 nucleus has a diameter about equal to the range of the nuclear force. In an Iron 56 nucleus every nucleon is attracting every other nucleon. If we go to a nucleus larger than Iron 56, then neighboring nucleons still attract each other, but protons on opposite sides of the nucleus now repel each other. This repulsion between distant protons leads to less binding energy per particle, and instability.

#### **NUCLEAR FISSION**

One way the instability of large nuclei shows up is in the process of nuclear fission, a process that is explained by the liquid drop model of the nucleus developed by Neils Bohr and John Wheeler in 1939.

In this model, we picture nuclear matter as being essentially an incompressible liquid. The nucleons cannot be pressed into each other, or pulled apart, but they are free to slide around each other like the water molecules in a drop of water. As a result of the liquid nature of nuclear matter, we can learn something about the behavior of nuclei by studying the behavior of drops of water.

In our discussion of entropy at the beginning of Chapter 18, we discussed a demonstration in which a stream of water is broken into a series of droplets by vibrating the hose leading to the stream. If you put a strobe light on the stream, you can stop the apparent motion of the individual droplets. The result is a strobe photograph of the projectile motion of the droplets.

If you use a closely focused television camera, you can follow the motion of individual drops. Adjust the strobe so that the drop appears to fall slowly, and you can watch an individual drop oscillate as it falls. As shown in Figure (1), the oscillation is from a rounded pancake shape (images 3 & 4) to a vertical jelly bean shape (images 6 & 7). Bohr and Wheeler proposed that similar oscillations should take place in a large nucleus like Uranium, particularly if

the nucleus were struck by some outside particle, like an errant neutron.

**Figure 1** Oscillations of a liquid drop.





Suppose we have an oscillating Uranium nucleus, and at the present time it has the dumbbell shape shown in Figure (2a,b). In this shape we have two nascent spheres (shown by the dotted circles) connected by a neck of nuclear matter. The nascent spheres are far enough apart that they are beyond the range  $R_0$  of the nuclear force, so that the electrical repulsion is stronger than the nuclear attraction. The only thing that holds this nucleus together is the neck of nuclear matter between the spheres.

If the Uranium nucleus is struck too vigorously, if the neck is stretched too far, the electric force will cause the two ends to fly apart, releasing a huge quantity of electrical potential energy. This process, shown in Figure (3) is called *nuclear fission*.

In the fission of Uranium 235, the large Uranium nucleus breaks up into two moderate sized nuclei, for example, Cesium 140 and Zirconium 94. Because larger nuclei have a higher percentage of neutrons than smaller ones, when Uranium breaks up into smaller,



Figure 2a Uranium nucleus in a dumbbell shape.



less neutron rich nuclei, some free neutrons are also emitted as indicated in Figure (3). These free neutrons may go out and strike other Uranium nuclei, causing further fission reactions.

If you have a small block of Uranium, and one of the Uranium nuclei fissions spontaneously (it happens once in a while), the extra free neutrons are likely to pass out through the edges of the block and nothing happens. If, however, the block is big enough, (if it exceeds a *critical mass* of about 13 pounds for a sphere), then neutrons from one fissioning nucleus are more likely to strike other Uranium nuclei than to escape. The result is that several other nuclei fission, and each of these cause several others to fission. Quickly you have a large number of fissioning nuclei in a process called a *chain reaction*. This is the process that occurs in an uncontrolled way in an atomic bomb and in a controlled way in a nuclear reactor.

The energy we get from nuclear fission, the energy from all commercial nuclear reactors, is electrical potential energy released when the two nuclear fragments fly apart. The fragments shown in Figure (3) are at that point well beyond the range  $R_0$  of the attractive nuclear force, and essentially feel only the repulsive electric force between the protons. These two balls of positive charge have a large positive electric potential energy which is converted to kinetic energy as the fragments fly apart.



**Figure 2b** Styrofoam model of a Uranium nucleus in a dumbbell shape.

To get a feeling for the amount of energy released in a fission reaction, let us calculate the electric potential energy of two fragments, say a Cesium and a Zirconium nucleus when separated by a distance  $2R_0$ , twice the range of the nuclear force.

In CGS units, the formula for the electric potential energy of 2 particles with charges  $Q_1$  and  $Q_2$  separated by a distance R is

$$\begin{array}{l} electric\\ potential\\ energy \end{array} \quad U_{electric} = \frac{Q_1 Q_2}{r} \quad CGS \ units \qquad (1) \end{array}$$

For our problem, let  $Q_1$  be the charge on a Cesium nucleus (55 protons) and  $Q_2$  the charge on a Zirconium nucleus (40 protons).

$$Q_{\text{Cesium}} = 55e$$
$$Q_{\text{Zirconium}} = 40e$$
$$r = 2R_0$$

and we get

$$U_{\text{electric}} = \frac{(55 \text{ e}) \times (40 \text{ e})}{2R_0}$$
$$= 1.1 \times 10^3 \frac{\text{e}^2}{R_0} \tag{2}$$

We would like to compare the energy released in nuclear fission reactions with the energies typically involved in chemical reactions. It takes a fairly violent chemical reaction to rip the electron completely out of a hydrogen atom. The amount of energy to do that, to ionize a hydrogen atom is  $e^2/r_b$  where  $r_b$  is the Bohr radius of  $5 \times 10^{-9}$  cm.

$$\left. \begin{array}{c} \text{energy to ionize} \\ \text{a hydrogen atom} \end{array} \right\} = \frac{e^2}{r_b} = 13.6 \text{ eV}$$
 (3)

We evaluated the number  $e^2/r_b$  earlier and found it to have a numerical value of 13.6 electron volts. This is a large amount of energy for a chemical reaction, more typical chemical reactions, arising from molecular forces, have involved energies in the 1 to 2 electron volt range. To compare the strength of nuclear fission reactions to chemical reactions, we can compare the electric potential energies in Equations 2 and 3. If we take the range  $R_0$  of the nuclear force to be 4 proton diameters then

$$R_0 = 4 \times (1.4 \times 10^{-13} \text{ cm}) = 5.6 \times 10^{-13} \text{ cm}$$

Since the Bohr radius is  $5 \times 10^{-9}$  cm, we see that  $R_0$  is essentially  $10^{-4}$   $R_b$  or ten thousand times smaller than the Bohr radius.

$$R_0 = 10^{-4} r_b \tag{4}$$

Substituting Equation 4 into 2 gives

$$U_{\text{electric}} = 1.1 \times 10^3 \frac{\text{e}^2}{\text{R}_0}$$
$$= 1.1 \times 10^3 \frac{\text{e}^2}{\left(10^{-4} \text{r}_b\right)}$$
$$= 1.1 \times 10^7 \left(\frac{\text{e}^2}{\text{r}_b}\right)$$

Using the fact that  $e^{2}/r_{b}$  has a magnitude of 13.6 electron volts, we get

$$U_{\text{electric}} = 1.1 \times 10^7 \times 13.6 \text{ eV}$$
  
=  $150 \times 10^6 \text{ eV}$  (5)  
=  $150 \text{ MeV}$ 

where 1MeV is one million electron volts. From Equation 5, we see that, per particle, some ten million times more electric potential energy is released in a nuclear fission reaction than in a violent chemical reaction. Many millions of electron volts are involved in nuclear reactions as compared to the few electron volts in chemical reactions. You can also see that a major reason for the huge amounts of energy in a nuclear reaction is the small size of the nucleus (the fact that  $R_0 \ll r_b$ ).

#### NEUTRONS AND THE WEAK INTERACTION

The stability of the iron nucleus and the instability of nuclei larger than Uranium results primarily from the fact that the attractive part of the nuclear force has a short range  $R_0$  over which it dominates the repulsive electric force between protons. The range  $R_0$  is about 4 proton diameters, the diameter of an iron nucleus. In larger nuclei, not all nucleons attract each other, and this leads to the kind of instability we see in a fissioning Uranium nucleus.

The range of the nuclear force is not the only important factor in determining the stability of nuclei. There is no electric repulsion between neutrons, neutrons are attracted equally to neutrons and protons. Adding neutrons to a nucleus increases the attractive nuclear force without enhancing the electric repulsion. This is why the most stable large nuclei have an excess of neutrons over protons. The neutron excess acts as a nuclear glue, diluting the repulsion of the protons.

If adding a few extra neutrons increases the stability of anucleus, why doesn't adding more neutrons give even more stable nuclei? Why is it that nuclei with too many excess neutrons are in fact unstable? Why can't we make nuclei out of pure neutrons and avoid the proton repulsion altogether?

The answer lies in the fact that, because of the weak interaction, and because of a small excess mass of a neutron, a neutron can decay into a proton and release energy. This is the beta decay reaction we discussed earlier, and is described by the equation

$$n^0 \rightarrow p^+ + e^- + \bar{\nu}^0 \tag{5}$$

The neutral neutron  $(n^0)$  decays into a positive proton  $(p^+)$ , a negative electron  $(e^-)$ , and a neutral antineutrino  $(\bar{v}^0)$ , thus electric charge is conserved in the process. It is called a beta  $(\beta)$  decay reaction because the electrons that come out were originally called *beta rays* before their identity as electrons was determined.

If you have an isolated free neutron, the ( $\beta$ ) decay of Equation 5 occurs with a half life of 15 minutes. Such a reaction can occur only if energy can be conserved in the process. But the neutron is sufficiently massive to decay into a proton and an electron and still have some energy left over. Expressing rest mass or rest energy in units of millions of electron volts (MeV), we have for the particles in the neutron decay reaction (5),

neutron rest mass  $m_n = 939.6 \text{ MeV}$ proton rest mass  $m_p = 938.3 \text{ MeV}$ electron rest mass  $m_e = 0.511 \text{ MeV}$  (6) neutrino rest mass  $m_v = 0$ 

where

 $1 \text{ MeV} = 10^6 \text{ eV} = 1.6 \times 10^{-6} \text{ ergs}$ 

You can see that the neutron rest mass is 1.3 MeV greater than that of a proton, and .8 MeV greater than the combined rest masses of the proton, electron and neutrino. Thus when a neutron ( $\beta$ ) decays, there is an excess of .8 MeV of energy that is released in the form of kinetic energy of the reaction products.

As we have mentioned, when the  $\beta$  decay process was first studied in the 1920s, the neutrino was unknown. What was observed was that in  $\beta$  decays, the proton and the electron carried out different amounts of energy, sometimes all of the available energy, but usually just part of it. To explain the missing energy, Wolfgang Pauli proposed the existence of an almost undetectable, uncharged, zero rest mass particle which Fermi named the little neutral one or neutrino. As bizarre as Pauli's hypothesis seemed at the time, it turned out to be correct. When a neutron decays, it decays into 3 particles, and the .8 MeV available for kinetic energy can be shared in various ways among the 3 particles. Because the neutrino has no rest mass, it is possible for the proton and electron to get all .8 MeV of kinetic energy. At other times the neutrino gets much of the kinetic energy, so that if you did not know about the neutrino, you would think energy was lost.

#### NUCLEAR STRUCTURE

Free neutrons decay in fifteen minutes, but neutrons inside a nucleus seem to live forever. Why don't they decay? The answer to this question is an energy balance. Like a rock dropped into a well, an atomic nucleus will fall down to the lowest energy state available. The neutron will decay if the result is a lower energy, less massive nucleus. Otherwise the neutron will be stable.

We have seen that an isolated neutron can decay into the less massive proton and release energy. Now consider a neutron in a nucleus. Take the simplest nucleus with a neutron in it, namely Deuterium. If that neutron decayed we would end up with a helium nucleus consisting of two protons only, plus an electron and a neutrino. We can write this reaction as

$${}^{2}\mathrm{H} \rightarrow {}^{2}\mathrm{He} + \mathrm{e}^{-} + \bar{\mathrm{v}} \qquad \begin{array}{c} deuterium \ decay \\ which \ does \ not \ happen \end{array} \tag{7}$$

$$\textcircled{PP} \longrightarrow \textcircled{PP} + \underbrace{\bullet}_{\mathrm{e}^{-}} + \underbrace{\bigvee}_{\mathrm{v}} \overset{(7)}{\mathrm{v}}$$

If the neutron in deuterium turns into a proton, the neutron sheds rest mass, but the resulting two proton nucleus has positive electric potential energy. We can estimate the amount of electric potential energy  $U_{pp}$  created by using the formula

electric potential  
energy of a 2  
proton nucleus 
$$U_{pp} = \frac{e^2}{2r_p}$$
 CGS units (8)

where  $r_p$  is the proton radius and  $2r_p = 2 \times 10^{-13}$  is the separation of the proton centers. The protons each have a charge + e. Putting numbers into Equation 8 gives

$$U_{p p} = \frac{e^{2}}{2r_{p}}$$

$$= \frac{\left(4.8 \times 10^{-10}\right)^{2}}{2 \times 10^{-13}} \text{ ergs}$$

$$= 1.15 \times 10^{-6} \text{ ergs}$$

$$= .72 \text{ MeV}$$
(9)

Our simple calculation shows that for the neutron to decay into a proton in deuterium, the neutron would have to create nearly as much electrical potential energy (.72MeV) as the available .8 MeV neutron mass energy. A more complete analysis shows that the available .8 MeV is not adequate for the neutron ( $\beta$ ) decay, with the result the neutron in deuterium is stable.

We can now see the competing processes involved in the formation of nuclei. To construct stable nuclei you want to add neutrons to give more attractive nuclear forces and dilute the repulsive electric forces between protons. However, the rest mass of a neutron is greater than the rest mass of a proton and an electron, and the weak interaction allows the neutron to decay into these particles. Thus neutrons can shed mass, and therefore energy, by decaying.

But when a neutron inside of a nucleus decays into a proton, it increases the electric potential energy of the nucleus. If the increase in the electric potential energy is greater than the mass energy released, as we nearly saw in the case of deuterium, then the neutron cannot decay.

The weak interaction is democratic, it allows a proton to decay into a neutron as well as a neutron to decay into a proton. The proton decay process is

$$p^+ \rightarrow n^0 + \bar{e}^+ + v^0$$
 inverse  $\beta$  decay (10)

where  $\bar{e}^+$  is the positively charged antielectron (positron). This *inverse*  $\beta$  *decay*, as it is sometimes called, does not occur for a free proton because energy is not conserved. The proton rest mass is less than that of a neutron, let alone that of a neutron and a positron combined.

However, if you construct a nucleus with too many protons, with too much electric potential energy, then the nucleus can get rid of some of its electric potential energy by converting a proton into a neutron. This may happen if enough electric potential energy is released to supply the extra rest mass of the neutron as well as the .5 MeV rest mass of the positron. We can now see the competing processes in an atomic nucleus. The weak interaction allows neutrons to turn into protons or protons into neutrons. But these  $\beta$  decay processes will occur only if energy can be released. Nuclei with too many neutrons have too much neutron mass energy, and can get rid of some of the mass energy by turning a neutron into a proton. Nuclei with too many protons have too much electrical potential energy, and can get rid of some of the electrical potential energy by turning a proton into a neutron. A stable nucleus is one that has neither an excess of mass energy nor electrical potential energy, a nucleus that cannot release energy either by turning protons into neutrons or vice versa.

To predict precisely which nuclei are stable and which are not requires a more detailed knowledge of the nuclear force than we have discussed here. But from what we have said, you can understand the general trend. For the light nuclei, the most stable are the ones with roughly equal numbers of protons and neutrons, nuclei with neither too much neutron mass energy or too much proton electrical potential energy. However, when nuclei become larger than the range of the attractive nuclear force, electric potential energy becomes more important . Excess neutrons, with the additional attractive nuclear binding force they provide, are needed to make the nucleus stable.

#### $\alpha$ (Alpha) Particles

In 1898 Ernest Rutherford, a young research student in Cambridge University, England, discovered that radioactive substances emitted two different kinds of rays which he named  $\alpha$  rays and  $\beta$  rays after the first two letters of the Greek alphabet. The negatively charged  $\beta$  rays turned out to be beams of electrons, and the positive  $\alpha$  rays were found to be beams of Helium 4 nuclei. Helium 4 nuclei, consisting of 2 protons and 2 neutrons, are thus also called  $\alpha$  particles. (Later Rutherford observed a third kind of radiation he called  $\gamma$  rays, which turned out to be high energy photons.) We have seen that  $\beta$  rays or electrons are emitted when a neutron sheds mass by decaying into a proton in a  $\beta$ decay reaction. But where do the  $\alpha$  particles come from?

A nucleus with an excess of electric potential energy can lose energy by converting one of its protons into a neutron in an inverse  $\beta$  decay reaction. This, however, is a relatively rare event. More commonly, the number of protons is reduced by ejecting an  $\alpha$  particle. Why the nucleus emits an entire  $\alpha$  particle or Helium 4 nucleus, instead of simply kicking out a single proton, is a consequence of an anomaly in the nuclear force. It turns out that a Helium 4 nucleus, with 2 neutrons and 2 protons, is an exceptionally stable, tightly bound object. If protons are to be ejected, they come out in pairs in this stable configuration rather than individually.

#### NUCLEAR BINDING ENERGIES

The best way to see the competition between the attractive nuclear force and the electric repulsive force inside atomic nuclei is to look at nuclear binding energies. Explicitly, we will look at the *binding energy per nucleon* for the most stable nuclei of each element. The binding energy per nucleon (proton or neutron) represents how much energy we would have to supply to pull the nucleus apart into separate free nucleons. The nuclear force tries to hold the nucleus together—make it more tightly bound—and therefore increases the binding energy. The electric force, which pushes the protons apart, decreases the binding energy.

You calculate the binding energy of a nucleus by subtracting the rest energy of the nucleus from the sum of the rest energies of the protons and neutrons that make up the nucleus. If you then divide by the number of nucleons, you get the binding energy per nucleon. We will go through an example of this calculation, and give you an opportunity to work out some yourself. Then we will look at a plot of these binding energies to see what the plot tells us.

#### Example 1

Given that a proton, a neutron, and a deuterium nucleus have the following rest energies, what is the binding energy per particle for the deuterium nucleus?

$m_{p}c^{2} = 938.3 \text{ MeV}$	proton rest energy
$m_{\rm n}c^2 = 939.6 {\rm MeV}$	neutron rest energy
$m_d c^2 = 1875.1 \text{ MeV}$	deuterium nucleus
	rest energy

#### Solution

Separately the proton and neutron in a deuterium nucleus have a total rest energy of

rest energy a separate proton and neutron = 1877.9 MeV Thus the total binding energy of the deuterium nucleus, the energy required to pull the particles apart is

total binding  
energy of the  
deuterium  
nucleus 
$$= \frac{1877.9 \text{ MeV}}{(\text{separate})} - \frac{1875.1 \text{ MeV}}{(\text{deuterium})}$$
  
 $= 2.8 \text{ MeV}$  (22)

Finally the binding energy per nucleon, there being 2 nucleons, is

$$\frac{\text{binding energy}}{\text{per nucleon}} = \frac{2.8 \text{ MeV}}{2 \text{ nucleons}}$$
$$= 1.4 \frac{\text{MeV}}{\text{nucleon}} \qquad (23)$$

#### Exercise 1

(20)

(21)

Given that the masses of the Helium 4 (2 protons, 2 neutrons), the Iron 56 (26 protons, 30 neutrons), and the Uranium 238 (92 protons, 146 neutrons) nuclei are

$$M_{\text{Helium 4}}c^2 = 3725.95 \text{MeV}$$

$$M_{\text{Iron 56}}c^2 = 52068.77 \text{MeV}$$

$$M_{\text{Uranium 238}}c^2 = 221596.94 \text{MeV}$$
(24)

find in each case, the binding energy per nucleon. Your results should be

$$\begin{array}{l} \text{bindingenergy} \\ \text{pernucleon} \end{array} \right\} = \begin{array}{l} 7.46 \, \text{MeV} \, (\text{Helium 4})) \\ = 9.20 \, \text{MeV} \, (\text{Iron 56}) \\ 8.02 \, \text{MeV} \, (\text{Uranium 238}) \end{array}$$

$$(25)$$

1

Figure (4) is a plot of the binding energy, per nucleon, of the most stable nuclei for each element. We have plotted increasing binding energy downward so that the plot would look like a well. The deeper down in the well a nucleus is, the more energy per particle that is required to pull the nucleus out of the well—to pull it apart. The deepest part of the well is at the Iron 56 nucleus, no other nucleus is more tightly bound. Moving down into the well represents a release of nuclear energy. There are two ways to do this. We can start with light nuclei and put them together (in a process called *nuclear fusion*), to form heavier nuclei, moving in and down from the left side in Figure (4). Or we can split apart heavy nuclei (In the process of nuclear fission), moving in and down from the right side. Fusion represents the release of nuclear force potential energy, while fission represents the release of electric force potential energy. When we get to the bottom, at Iron 56, there is no energy to be released either by fusion or fission.



#### Figure 4

The nuclear energy well. The graph shows the amount of energy, per nucleon, required to pull the nucleus apart into separate neutrons and protons.

20-11

The reason Iron 56 is at the bottom of the well is because the diameter of an iron nucleus is about equal to the range of the nuclear force. As you build up to Iron 56, adding more nucleons increases the number of attractive forces between particles, and therefore increases the strength of the binding. At Iron 56, you have the largest nucleus in which every particle attracts every other particle. The diameter of the Iron 56 nucleus is the distance over which the attractive nuclear force is stronger than the repulsive electric force.

When you build nuclei larger than Iron 56, the protons on opposite sides of the nucleus are far enough apart that the electric force is stronger and the particles repel. Now, adding more particles reduces the binding energy per particle and produces less stable nuclei. When a nucleus become as large as uranium, the impact of a single neutron can cause the nucleus to split apart into two smaller, more stable nuclei in the nuclear fission process.

There are some bumps in the graph of nuclear binding energies, bumps representing details in the structure of the nuclear force. The most striking anomaly is the Helium 4 nucleus which is far more tightly bound than neighboring nuclei. This tight binding of Helium 4 is the reason, as we mentioned, that  $\alpha$  particles (Helium 4 nuclei) rather than individual protons are emitted in radioactive decays. But overlooking the bumps, we have the general feature that nuclear binding energies increase up to Iron 56 and then decrease thereafter. The importance of knowing the nuclear binding energy per nucleon is that it tells us whether energy will be released in a particular nuclear reaction. If the somewhat weakly bound uranium nucleus (7.41 MeV/ nucleon) splits into two more tightly bound nuclei like cesium (8.16 MeV/nucleon) and zirconium (8.41 MeV/ nucleon), energy is released. At the other end of the graph, if we combine two weakly bound deuterium nuclei (2.8 MeV/nucleon) to form a more tightly bound Helium 4 nucleus (7.1 MeV/nucleon), energy is also released. Any reaction that moves us toward the Iron 56 nucleus releases energy. On the small nucleus side we get a release of energy by combining small nuclei to form bigger nuclei. But once past Iron 56, we get a release of energy by splitting nuclei apart for form smaller ones.

#### **NUCLEAR FUSION**

The process of combining small nuclei to form larger ones is called nuclear fusion. From our graph of nuclear binding energies in Figure (4), we see that nuclear fusion releases energy if the resulting nucleus is smaller than Iron 56, but costs energy if the resulting nucleus is larger. This fact has enormous significance in the life of stars and the formation of the elements.

Most stars are created from a gas cloud rich in hydrogen gas. When the cloud condenses, gravitational potential energy is released and the gas heats up. If the condensing cloud is massive enough, if the temperature becomes hot enough, the hydrogen nuclei begin to fuse. After several reactions they produce Helium 4 nuclei, releasing energy in each reaction. The fusion of hydrogen to form helium becomes the source of energy for the star for many years to come.

Unlike fission, fusion requires high temperatures in order to take place. Consider the reaction in which two hydrogen nuclei (protons) fuse to produce a deuterium nucleus plus a positron and a neutrino. (When the two protons fuse, the resulting nucleus immediately gets rid of its electrical potential energy by having one proton turn into a neutron in an inverse  $\beta$  decay process.)

The fusion of the two protons will take place if the protons get closer together than the range of the nuclear force about—4 proton diameters. Before they get that close they repel electrically. Only if the protons were initially moving fast enough, were hot enough, can they get close enough to get past the electrical repulsion in order to feel the nuclear attraction.

A good way to picture the situation is to think of yourself as sitting on one of the protons, and draw a graph of the potential energy of the approaching proton, as shown in Figure (5). When the proton separation r is greater than the range  $R_0$  of the nuclear force, the protons repel and the incoming proton has to climb a potential energy hill. At  $R = R_0$ , the net force turns attractive and the potential energy begins to decrease, forming a deep well when the particles are near to touching. Energy is released when the incoming proton falls into the well, but the incoming proton must have enough kinetic energy to get over the electrical potential energy barrier before the fusion can take place.

Using our formula  $Q_1Q_2/r$  for electric potential energy, we can make a rough estimate of the kinetic energy and the temperature required for fusion. Consider the fusion of two protons where  $Q_1 = Q_2 = e$ . For the protons to get within a distance  $R_0$ , the incoming proton must climb a barrier of height

electric potential  
energy of 2 protons  
a distance R<sub>0</sub> apart 
$$= \frac{e^2}{R_0}$$
$$= \frac{\left(4.8 \times 10^{-10}\right)^2}{4 \times 1.4 \times 10^{-13}} \quad (26)$$
$$= 4.1 \times 10^{-7} \text{ ergs}$$

This number,  $e^2/R_0 = 4.1 \times 10^{-7}$  ergs , is the amount of kinetic energy an incoming proton must originally have in order to get within a distance of approximately  $R_0$  of another proton. Only when it gets within this distance can the nuclear force take over and fusion take place.



Figure 5

When you shove two protons together, you first have to overcome the electric repulsion before the nuclear attraction dominates.

In our earlier discussion of temperature, we saw that the average kinetic energy of a particle in a gas of temperature T was 3/2 kT. If we had a gas of hydrogen so hot that the average proton could enter into a fusion reaction, the average kinetic energy of the protons would have to be  $4.1 \times 10^{-7}$  ergs. The temperature T<sub>f</sub> at which this would happen is found by equating  $4.1 \times 10^{-7}$  ergs to 3/2 kT to give

$$\frac{3}{2}kT = 4.1 \times 10^{-7} \text{ ergs}$$

$$T = \frac{2}{3} \frac{4.1 \times 10^{-7} \text{ ergs}}{1.38 \times 10^{-16} \frac{\text{ ergs}}{\text{ kelvin}}}$$

$$T = 2 \times 10^{9} \text{ kelvin} \qquad (27)$$

This temperature, two billion degrees kelvin, is a huge overestimate. At this temperature, the *average* proton in the gas would enter into a fusion reaction. If we heated a container of hydrogen to this temperature, the entire collection of protons would fuse after only a few collisions, and the fusion energy would be released almost instantaneously. We would have what is known as a *hydrogen bomb*.

In a star, the fusion of hydrogen takes place at the much lower temperatures of about **20 million** degrees. At 20 million degrees, only a small fraction of the protons have enough kinetic energy to enter into a fusion reaction. At these temperatures the hydrogen is consumed at a slow steady rate in what is known as a **controlled fusion reaction**.

#### **STELLAR EVOLUTION**

The story of the evolution of stars provides an ideal setting to illustrate the interplay of the four basic interactions. In this chapter so far we have been focusing on the basic consequences of the interplay of the nuclear, electric, and weak interactions at the level of atomic nucleus. Add gravity and you have the story of stellar evolution.

A star is born from a cloud of gas, typically rich in hydrogen, that begins to collapse gravitationally. As the cloud collapses, *gravitational potential energy* is released which heats the gas. If the temperature does not get hot enough to start the fusion of the hydrogen nuclei, that's more or less the end of the story and you have a *proto star*, something around the size of the planet Jupiter or smaller.

If there is more mass in the collapsing gas cloud, more gravitational potential energy will be released, and the temperature will rise enough to start the fusion of hydrogen. How hot the center of the star becomes depends on the mass of the star. In a star, like our sun for example, there is a balance between the gravitational attraction and the thermal pressure. The greater the mass, the greater the gravitational attraction, and the stronger the thermal pressure must be.

In our discussion of pressure in chapter 17, we saw that when a balloon was cooled by liquid nitrogen, removing the thermal energy and pressure of the air molecules inside, the balloon collapsed. (Figures 17-19.) Similar processes occur in a star, except that the confining force of the rubber is replaced by the confining force of gravity. The sun is a ball of hot gas. Gravity is trying to squeeze the gas inward, and the thermal pressure of the gas prevents it from doing so. There is a precise balance between the thermal pressure and gravity.



**Figure 17-19** Balloon collapsing in liquid nitrogen.

From the earth the sun looks more like a solid object than a ball of gas. The sun has a definite edge, an obvious surface with spots and speckles on it. The appearance of a sharp surface is the result of the change in temperature of the gas with height. The hottest part of the sun or any star is the center. Here the gas is so hot, the thermal collisions are so violent, that the electrons are knocked out of the hydrogen atoms and all the hydrogen is ionized. The gas is what is called a *plasma*. An ionized gas or plasma is opaque, light is absorbed by the separate charged particles.

As you go out from the center of the star, the temperature drops. When you go out far enough, when the temperature drops to about 3000 kelvins, the electrons recombine with the nuclei, you get neutral atoms, and the gas becomes transparent. This transition from an opaque to transparent gas occurs rather abruptly, giving us what we think of as the surface of the sun.

Returning to the balance of gravitational attraction and thermal pressure, you can see that the more mass in the star the stronger gravity is, and the greater the thermal pressure required to balance gravity. To increase the thermal pressure, you have to increase the temperature. Thus the more massive a star, the hotter it has to be.

The proton fusion reaction we have discussed is well suited for supplying any required temperature. We have seen that the rate at which fusion takes place depends very much on the temperature. At 20 million degrees, only a small fraction of the protons have enough thermal kinetic energy to fuse. At 2 billion degrees, the average proton has enough energy to fuse, and any hydrogen at this temperature would burn immediately. There is a range of burning rates between these two extremes. As a result, with increasing temperature the fusion reaction goes faster, supplies energy at a greater rate, and maintains the higher temperature. Thus when a new star forms it collapses until there is a balance between the gravitational force and thermal pressure. Whatever thermal pressure is required is supplied by the heat generated by the fusion reaction. The more thermal pressure needed, the higher the temperature required and the faster the hydrogen burns.

One often refers to the region near the center of the star where the fusion reaction is taking place as the *core* of the star. Our sun, with a temperature in the core of 20 million degrees, is burning hydrogen at such a rate that the hydrogen supply will last 10 billion years. Since the sun is 5 billion years old, about half the available hydrogen in the core is used up.

A more massive star, like the star that blew up to give us the 1987 supernova event, burns its hydrogen at a much shorter time. That star was about 18 times as massive as the sun, about 40,000 times as bright, and burned its hydrogen in its core so fast that the hydrogen lasted only about 10 million years.

The difference between different mass stars shows up most dramatically after the hydrogen fuel is used up. What will happen to our sun is relatively calm. When our sun uses up the hydrogen, the core will start to cool and collapse. But the collapse releases large amounts of gravitational potential energy that heats the core to higher temperatures than before. This hotter core becomes very bright, so bright that the light from the core, when it works its way out to the surface, exerts a strong radiation pressure on the gas at the surface. This radiation pressure will cause the surface of the sun to expand until the diameter of the sun is about equal to the diameter of the earth's orbit. At this point the sun will have become what is called a *red giant star*, with the earth orbiting slightly inside. It is not a very pleasant picture for the earth, but it will not happen for another five billion years.

Once the sun, as a red giant star, radiates the energy it got from the gravitational collapse, it will gradually cool and collapse until the atoms push against each other. It will be the electric force between the atomic electrons that will halt the gravitational collapse of the sun. At that point the sun will become a ball of highly compressed atomic matter about the size of the earth. Initially it will be quite bright, an object called a *white dwarf* star, but eventually it will cool and darken.

The story was very different for the star that gave rise to the supernova explosion. That star, with its mass of about 18 times that of the sun, burned its hydrogen in 10 million years. At that point the star had a core, about 30 per cent of the star, consisting mostly of Helium 4, the tightly bound nucleus that is the end result of hydrogen fusion. Computer simulations tell us that for the next tens of thousands of years, the helium core was compressed from a density of 6 to 1,100 grams per cubic centimeter, and the temperature rose from 40 million to 190 million degrees kelvin.

The temperature of 190 million degrees is high enough to cause helium nuclei to fuse, forming carbon and oxygen. Higher temperatures are required to fuse helium nuclei, because each helium nucleus has two protons and a charge + 2e. Thus the electric potential barrier  $Q_1Q_2/R_0$  is four times as high as it is for proton fusion, and the helium nuclei need four times as much kinetic energy to fuse.

At these higher temperatures the core radiated more light, causing the outer layers of the star, mostly unburned hydrogen, to expand to about twice the size of the earth's orbit. It had become a *red supergiant*.

The helium in the core lasted less than a million years, leaving behind a collapsing core of carbon and oxygen. The temperature rose to about 740 million degrees where the carbon ignited to form neon, magnesium and sodium. When the carbon was used up in about 12,000 years, further collapse raised the temperature to 1.6 billion degrees where neon ignited.

In successively shorter times at successively higher temperatures the more massive elements were created and burned. After neon, there was carbon, then oxygen at 2.1 billion degrees, and finally silicon and sulfur at 3.4 billion degrees. The neon burned in about 12 years, the oxygen in 4 years, and the silicon in just a week.

One of the reasons for the accelerated pace of burning at the end is that, at temperatures over half a billion degrees, the star has a more efficient way of getting rid of energy than emitting light. At these temperatures some of the photons are energetic enough to create electron-positron pairs which usually annihilated back into photons but *sometimes into neutrinos*. Whereas photons take thousands of years to carry energy from the core of a star to the surface, neutrinos escape immediately. Thus when the star reached half a billion degrees, it sprung a neutrino heat leak, and the collapse and burning went much faster.

The successive stages of burning took place in smaller and smaller cores, leaving shells of unburned elements. Unburned hydrogen filled the outer volume of the star. Inside was a shell of unburned helium and inside that successive shells of unburned carbon, oxygen, then a mixture of neon, silicon and sulfur, and a shell of silicon and sulfur.

In the center was iron. Iron was what resulted when the silicon and sulfur burned. And iron is the end of the road. As we have seen in Figure (4) iron is the most stable atomic nucleus. Energy is *released* when you fuse nuclei to form a nucleus smaller than iron, but it *costs* energy to create nuclei larger than iron. The iron core of the star was dead ash not a fuel.

By the time the star had an iron core, it was rapidly radiating energy in the form of neutrinos, but had run out of fuel. At this point the iron core, which had a mass of about 1.4 times the mass of the sun, began to collapse due to the lack of support by thermal pressure. When the sun runs out of energy, its collapse will be halted by the electric repulsion between atomic electrons. In the 1987 supernova star, the gravitational forces were so great that the electrons were essentially crammed back into the nuclei, the protons converted to neutrons, and the core collapsed into a ball of neutrons about 100 miles in diameter.

This collapse took a few tenths of a second, and created a shock wave that rapidly spread to the outer layers of the star. Vast quantities of neutrinos were created in the collapse, and escaped over the next 10 or so seconds. The shock wave reached the surface of the star 3 hours later, blowing off the surface of the star and starting a burst of light 3 hours behind the burst of neutrinos. The light and neutrinos raced each other for 180,000 years, and the neutrinos were still at least 2 hours ahead when they got to the earth.

When a supernova explodes, the outer shells of hydrogen, helium, carbon, neon, magnesium, sodium, silicon, sulfur and iron are blown out to form a new dust cloud. Such a dust cloud—the Crab Nebula, produced by the 1054 supernova explosion—is shown in Figure (6). From this cloud new stars and planets will form, stars and planets rich in the heavier elements created in the star and recycled into space by the supernova explosion.

Elements heavier than iron are also in the supernova remnants. So much energy is released in the collapsing core of the supernova that elements heavier than iron are created by fusion, even though this fusion costs energy. All the silver and gold in your watchband and ring, the iodine in your medicine cabinet, the mercury in your thermometer, and lead in your fishing sinker, all of these elements which lie beyond iron, were created in the flash of a supernova explosion. Without supernova explosions, the only raw materials for the formation of stars and planets would be hydrogen, some deuterium, and a trace of other light elements left over from the Big Bang that created the universe. How elements are formed in stars has been a fascinating detective story carried out over the past 40 years. The pioneering work, carried out by William Fowler, Fred Hoyle, and others, involved a careful study of nuclear reactions in the laboratory and then a modeling of how stars should evolve based on the known reactions. On one occasion a nuclear reaction was predicted to exist because it had to be there for stars to evolve. The reaction was then found in the laboratory, exactly as predicted.

The modeling of stellar evolution, using experimental data on nuclear reactions, and large computer programs, has quite successfully predicted the relative abundance of the various elements, as well as many features of stellar evolution such as the expansion of a star into a red giant when the hydrogen fuel in its core is used up. The success of these models leads us to believe that the details we described about what happened in the very core of a star about to explode, actually happened as described. An exciting consequence of the 1987 supernova explosion was that we got a glimpse into the core of the exploding star, a glimpse provided by the neutrinos that took 10 seconds, as predicted, to escape from the core. The neutrinos also arrived three hours before the photons, as predicted by the computer models.



#### Figure 6a

The Crab Nebula. The arrow points to the pulsar that created the nebula. (Above photo Hale Observatories. 6b: 1950 Photograph by Walter Baade, 1964 by Güigo Munch, composite by Munch and Virginia Trimble.)

#### **NEUTRON STARS**

One of the great predictions of astronomy based on the physical properties of matter was made by a young physicist/astronomer S. Chandrasekhar in the 1930s. With some relatively straightforward calculations, Chandrasekhar predicted that if a cooling, collapsing star had a mass greater than 1.4 times the mass of the sun, the force of gravity would be strong enough to cram the atomic electrons down into the nuclei, converting the protons to neutrons, leaving behind a ball of neutrons about 10 miles in diameter.

Chandrasekhar talked about this idea with his sponsor Sir Arthur Eddington, who was, at the time, one of the most famous astronomers in the world. In private, Eddington agreed with Chandrasekhar's calculations, but when asked about them in the 1932 meeting of the Royal Astronomical Society, Eddington replied that he did not believe that such a process could possibly occur. Chandrasekhar's ideas were dismissed by the astronomical community, and a discouraged Chandrasekhar left astronomy and went into the field of hydrodynamics and plasma physics where he made significant contributions. In 1967, the graduate student Jocelyn Bell, using equipment devised by Antony Hewish, observed an object emitting extremely sharp radio pulses that were 1.337 seconds apart. By the end of the year, up to ten such pulsing objects were detected, one with pulses only 89 milliseconds apart. After eliminating the possibility that the radio pulse was communication from an advanced civilization, it was determined that the signals were most likely from an objects rotating at high speeds. A star cannot rotate that fast unless it is very compact, less than 100 miles in diameter. The only candidate for such an object was Chandresekhar's neutron star.

Many other pulsing stars—*pulsars*—have been discovered. The closest sits at the center of the explosion that created the Crab Nebula seen in Figure (6a). A superposition of photographs taken in 1950 and 1964, Figure (6b), shows that the gas in the Crab Nebula is expanding away from the star marked with an arrow. Taking a high speed moving picture of this star, something that one does not usually do when photographing stars, shows that this star turns on and off **33 times a second** as seen in Figure (7). This is the neutron star left behind when the supernova exploded.



#### Figure 6b

Expansion of the Crab Nebula. Two photographs, taken 14 years apart, the first printed in white, the second dark, show the expansion centered on the pulsar.



**Figure 7** *Neutron star in Crab Nebula, turns on and off 33 times/sec. (Exposure from the Lick Observatory.)* 

We have now been able to study many pulsars, and know that the typical neutron star is a ball of neutrons about 10 miles in diameter, rotating at rates up to nearly 1000 revolutions per second! We can detect neutron stars because they have a bright spot that emits a beam of radiation. We see the pulses of radiation when the beam sweeps over us much as the captain of a ship sees the bright flash from a lighthouse when the beam sweeps past.

Computer models suggest that the bright spot is created by the magnetic field of the star, a field that was tied to the material in the core of the star and was strengthened as the core collapsed. Charged particles in the 'atmosphere' of the neutron star spiral around the magnetic field lines striking the star at the magnetic poles. On the earth, charged particles spiraling around the earth's magnetic field lines strike the earth's atmosphere at the magnetic poles, creating the aurora borealis and aurora australis, the northern and southern lights that light up extreme northern and southern night skies. On a neutron star, the aurora is much brighter, also the atmosphere much thinner, only a few centimeters thick. (In Chapter 28 on Magnetism, we will talk about the motion of charged particles in a magnetic field.)

#### NEUTRON STARS AND BLACK HOLES

As predicted by Chandrasekhar, when a cooling star is more than 1.4 times as massive as the sun, the gravitational attraction becomes strong enough to overcome the electronic structure of matter, shoving the electrons into the nuclei and leaving behind a ball of neutrons. A neutron star is essentially a gigantic nucleus in which the attractive gravitational force which holds the ball together, is balanced by the repulsive component of the nuclear force which keeps the neutrons from squeezing into each other.

In our discussion of atomic nuclei, it was the attractive component of the nuclear force that was of the most interest. It was the attractive part that overcame the electric repulsion between protons. Now in the neutron star, gravity is doing the attracting and the nuclear force is doing the repelling.

Einstein's special theory of relativity sets a limit on how strong the repulsive part of the nuclear force can be. We can see why with the following qualitative arguments.

The harder it is to shove two nucleons into each other, the stronger the repulsive part of the nuclear force, the more incompressible nuclear matter is. Now in our beginning discussions of the principle of relativity in Chapter 1, we saw that the speed of a sound wave depended upon the compressibility of the material through which the sound was moving. We used a stretched Slinky for our initial demonstrations of wave motion because a stretched Slinky is very easy to compress, with the result that Slinky waves move very slowly, about 1 foot per second.

Air is much more incompressible than a Slinky (try blowing air into a Coke bottle), with the result that sound waves in air travel about 1000 times faster than slinky waves. Water is more incompressible yet, and sound travels through water 5 times faster than in air. Because steel is even more incompressible than water, sound travels even faster in steel, about 4 times faster than in water. The most incompressible substance known is nuclear matter. It is so incompressible, the repulsive force between nucleons is so great that the calculated speed of sound in nuclear matter approaches the speed of light. And that's the limit. Nothing can be so rigid or incompressible that the speed of sound in the substance exceeds the speed of light. This fact alone tells us that there is a limit to how rigid matter can be, how strong repulsive forces can become. The repulsive part of the nuclear force approaches that limit.

There is, however, no limit to the strength of attractive forces. The attractive gravitational force in a neutron star simply depends upon the amount of mass in the star. The more mass, the stronger the force.

From this we can conclude that we are in serious trouble if the neutron star gets too big. It is estimated that in a neutron star with a mass 4 to 6 times the mass of the sun, the attractive gravitational force will exceed the repulsive component of the nuclear force, and the neutrons will begin to collapse into each other.

As the star starts to collapse, gravity gets still stronger. But gravity has just crushed the strongest known repulsive force. At some point during the collapse, gravity will become strong enough to crush any possible repulsive force. According to the laws of physics, *as we know them*, nothing can stop the further collapse of the star, perhaps down to a mathematical point, or at least down to a size so small that new laws of physics take over.

The problem with finding black holes is that, as their name suggests, they do not emit light. The only way we have of detecting black holes is by their gravitational effect on other objects.

It turns out that a good fraction of the stars in the universe come in binary pairs. Having a pair of stars form is an effective way of taking up the angular momentum of a collapsing gas cloud that was initially rotating. If Jupiter had been just a bit bigger, igniting its own nuclear reactions, then the earth would have been located in a binary star system. If one of a pair of binary stars is a black hole, two detectable effects can occur. If the stars are in close orbit, the black hole will suck off the outer layers of gas from the visible companion, as indicated in the artist's conception, Figure (12). According to computer models, the gas that is drawn off from the companion star goes into orbit around the black hole, forming what is called an accretion disk around the black hole. The gas in the accretion disk is moving very rapidly, at speeds approaching the speed of light. As a result of the high speeds, and turbulence in the flow, the particles in the accretion disk emit vast quantities of X rays as they spiral down toward the black hole. Strong X ray emission is thus a signature that an object may be a binary star system with one of the stars being a black hole.



#### Figure 8

Painting of the gas from a blue giant star being sucked into a black hole. (From the May, 1974 National Geographic, Artist Victor J. Kelley.) Since X rays can be emitted in other ways, a further check is needed to be sure that a black hole is involved. By studying the orbit of the visible companion, one can determine the mass of the invisible one (essentially using Kepler's third law, in a slightly modified form). The test of whether the invisible companion is a black hole is whether its mass is over 6 solar masses. If it is, then no dark object could withstand the gravitational forces involved. The first candidate for an object fitting this description is the X ray source in the constellation of Cygnus, an object known as *Cygnus X1*.

Rather than being scarce, hard-to-find objects, black holes may play a significant role in the structure of the universe. There is good evidence, from the study of the motions of stars, that a gigantic black hole, with a mass of millions of solar masses, may lie at the heart of our galaxy and other galaxies as well. And recent studies have indicated that there may be a black hole at the center of globular clusters. Wherever black holes may be, whatever their role in our universe, one fact stands out—gravity, a force too weak even to be detected on an atomic scale, can under the right circumstances become the strongest force of all, strong enough to crush matter out of existence.

# Physics 2000



E. R. Huggins **Dartmouth College** 

### Part 2 E & M, **Quantum Mechanics**, **Optics, Calculus**



physics2000.com

## Chapter 23 Fluid Dynamics

Since the earth is covered by two fluids, air and water, much of our life is spent dealing with the dynamic behavior of fluids. This is particularly true of the atmosphere where the weather patterns are governed by the interaction of large and small vortex systems, that sometime strengthen into fierce systems like tornados and hurricanes. On a smaller scale our knowledge of some basic principles of fluid dynamics allows us to build airplanes that fly and sailboats that sail into the wind.

In this chapter we will discuss only a few of the basic concepts of fluid dynamics, the concept of the velocity field, of streamlines, Bernoulli's equation, and the basic structure of a well-formed vortex. While these topics are interesting in their own right, the subject is being discussed here to lay the foundation for many of the concepts that we will use in our discussion of electric and magnetic phenomena. This chapter is fairly easy reading, but it contains essential material for our later work. It is not optional.

#### The Current State of Fluid Dynamics

The ideas that we will discuss here were discovered well over a century ago. They are simple ideas that provide very good predictions in certain restricted circumstances. In general, fluid flows can become very complicated with the appearance of turbulent motion. Only in the twentieth century have we begun to gain confidence that we have the correct equations to explain fluid motion. Solving these equations is another matter and one of the most active research topics in modern science. Fluid theory has been the test bed of the capability of modern super computers as well as the focus of attention of many theorists. Only a few years ago, from the work of Lorenz it was discovered that it was not possible, even in principle, to make accurate long-range forecasts of the behavior of fluid systems, that when you try to predict too far into the future, the chaotic behavior of the system destroys the accuracy of the prediction.

Relative to the current work on fluid behavior, we will just barely touch the edges of the theory. But even there we find important basic concepts such as a vector field, streamlines, and voltage, that will be important throughout the remainder of the course. We are introducing these concepts in the context of fluid motion because it is much easier to visualize the behavior of a fluid than some of the more exotic fields we will discuss later.

#### THE VELOCITY FIELD

Imagine that you are standing on a bridge over a river looking down at the water flowing underneath you. If it is a shallow stream the flow may be around boulders and logs, and be marked by the motion of fallen leaves and specks of foam. In a deep, wide river, the flow could be quite smooth, marked only by the eddies that trail off from the bridge abutments or the whipping back and forth of small buoys.

Although the motion of the fluid is often hard to see directly, the moving leaves and eddies tell you that the motion is there, and you know that if you stepped into the river, you would be carried along with the water.

Our first step in constructing a theory of fluid motion is to describe the motion. At every point in the fluid, we can think of a small "particle" of fluid moving with a velocity  $\vec{v}$ . We have to be a bit careful here. If we picture too small a "particle of fluid", we begin to see individual atoms and the random motion between atoms. This is too small. On the other hand, if we think of too big a "particle", it may have small fluid eddies inside it and we can't decide which way this little piece of fluid is moving. Here we introduce a not completely justified assumption, namely that there is a scale of distance, a size of our particle of fluid, where atomic motions are too small to be seen and any eddies in the fluid are big enough to carry the entire particle with it. With this idealization, we will say that the velocity  $\vec{v}$  of the fluid at some point is equal to the velocity of the particle of fluid that is located at that point.

We have just introduced a new concept which we will call the "velocity field". At every point in a fluid we define a vector  $\vec{v}$  which is the velocity vector of the fluid particle at that point. To formalize the notation a bit, consider the point labeled by the coordinates (x, y, z). Then the velocity of the fluid at that point is given by the vector  $\vec{v}$  (x, y, z), where  $\vec{v}$  (x, y, z) changes as we go from one point to another, from one fluid particle to another. As an example of what we will call a *velocity field*, consider the bathtub vortex shown in Figure (1a). From the top view the water is going in a nearly circular motion around the vortex core as it spirals down the funnel. We have chosen Points A, B, C and D, and at each of the points drawn a velocity vector to represent the velocity of the fluid particle at that point. The velocity vectors are tangent to the circular path of the fluid and vary in size depending on the speed of the fluid. In a typical vortex the fluid near the core of the vortex moves faster than the fluid out near the edge. This is represented in Figure (1b) by the fact that the vector at Point D, in near the core, is much longer than the one at Point A, out near the edge.



Figure 1a The "bathtuh" vort

The "bathtub" vortex is easily seen by filling a glass funnel with water, stirring the water, and letting the water flow out of the bottom.



#### Figure 1b

Looking down from the top, we see the water moving around in a circular path, with the water near the core moving faster. The velocity vectors, drawn at four different points, get longer as we approach the core.

#### **The Vector Field**

The velocity field, illustrated in Figure (1) is our first example of a more general concept called a *vector field*. The idea of a vector field is simply that at every point in space there is a vector with an explicit direction and magnitude. In the case of the velocity field, the vector is the velocity vector of the fluid particle at that point. The vector  $\vec{v}$  (x, y, z) points in the direction of motion of the fluid, and has a magnitude equal to the speed of the fluid.

It is not hard to construct other examples of vector fields. Suppose you took a 1 kg mass hung on the end of a spring, and carried it around to different parts of the earth. At every point on the surface where you stopped and measured the gravitational force  $\vec{F} = m\vec{g} = \vec{g}$  (for m = 1) you would obtain a force vector that points nearly toward the center of the earth, and has a magnitude of about 9.8 m/sec<sup>2</sup> as illustrated in Figure (2). If you were ambitious and went down into tunnels, or up on very tall buildings, the vectors would still point toward the center of the earth, but the magnitude would



Figure 2

We can begin to draw a picture of the earth's gravitational field by carrying a one kilogram mass around to various points on the surface of the earth and drawing the vector  $\vec{g}$  representing the force on that unit mass (m = 1) object.

vary a bit depending how far down or up you went. (Theoretically the magnitude of  $\vec{g}$  would drop to zero at the center of the earth, and drop off as  $1/r^2$  as we went out away from the earth). This quantity  $\vec{g}$  has a magnitude and direction at every point, and therefore qualifies as a vector field. This particular vector field is called the *gravitational field* of the earth.

It is easy to describe how to construct the gravitational field  $\vec{g}$  at every point. Just measure the magnitude and direction of the gravitational force on a non-accelerated 1 kg mass at every point. What is not so easy is to picture the result. One problem is drawing all these vectors. In Figure (2) we drew only about five  $\vec{g}$  vectors. What would we do if we had several million measurements?

The gravitational field is a fairly abstract concept—the result of a series of specific measurements. You have never seen a gravitational field, and at this point you have very little intuition about how gravitational fields behave (do they "behave"? do they do things?). Later we will see that they do.

In contrast you have seen fluid motion all your life, and you have already acquired an extensive intuition about the behavior of the velocity field of a fluid. We wish to build on this intuition and develop some of the mathematical tools that are effective in describing fluid motion. Once you see how these mathematical tools apply to an easily visualized vector field like the velocity field of a fluid, we will apply these tools to more abstract concepts like the gravitational field we just mentioned, or more importantly to the electric field, which is the subject of the next nine chapters.


(a) Edge-view of the so called Hele-Shaw cell



(b) Flow around a circular object.



(c) Flow around airplane wing shapes.

#### Figure 3

In a Hele-Shaw cell, bands of water and ink flow down through a narrow gap between sheets of glass. With this you can observe the flow around different shaped objects placed in the gap. The alternate black and clear bands of water and ink mark the streamlines of the flow.

## **Streamlines**

We have already mentioned one problem with vector fields—how do you draw or represent so many vectors? A partial answer is through the concept of streamlines illustrated in Figure (3). In that figure we have two plates of glass separated by a narrow gap with water flowing down through the gap. In order to see the path taken by the flowing water, there are two fluid reservoirs at the top, one containing ink and the other clear water. The ink and water are fed into the gap in alternate bands producing the streaks that we see. Inside the gap are a plastic cut-out of both a cylinder and a cross section of an airplane wing, so that we can visualize how the fluid flows past these obstacles.

The lines drawn by the alternate bands of clear and dark water are called streamlines. Each band forms a separate stream, the clear water staying in clear streams and the inky water in dark streams. What these streams or streamlines tell us is the direction of motion of the fluid. Because the streams do not cross and because the dark fluid does not mix with the light fluid, we know that the fluid is moving along the streamlines, not perpendicular to them. In Figure (4) we have sketched a pair of streamlines and drawn the velocity vectors  $\vec{v}_1$ ,  $\vec{v}_2$ ,  $\vec{v}_3$  and  $\vec{v}_4$  at four points along one of the streams. What is obvious is that the velocity vector at some point must be parallel to the streamline at that point, for that is the way the fluid is flowing. The streamlines give us a map of the directions of the fluid flow at the various points in the fluid.

## Figure 4

Velocity vectors in a streamline. Since the fluid is flowing along the stream, the velocity vectors are parallel to the streamlines. Where the streamlines are close together and the stream becomes narrow, the fluid must flow faster and the velocity vectors are longer.



#### **Continuity Equation**

When we have a set of streamlines such as that in Figure (4), we have a good idea of the directions of flow. We can draw the direction of the velocity vector at any point by constructing a vector parallel to the streamline passing through that point. If the streamline we have drawn or photographed does not pass exactly through that point, then we can do a fairly good job of estimating the direction from the neighboring streamlines.

But what about the speed of the fluid? Every vector has both a magnitude and a direction. So far, the streamlines have told us only the directions of the velocity vectors. Can we determine or estimate the fluid speed at each point so that we can complete our description of the velocity field?

When there is construction on an interstate highway and the road is narrowed from two lanes to one, the traffic tends to go slowly through the construction. This makes sense for traffic safety, but it is just the wrong way to handle an efficient fluid flow. The traffic should go faster through the construction to make up for the reduced width of the road. (Can you imagine the person with an orange vest holding a sign that says "Fast"?) Water, when it flows down a tube with a constriction, travels faster through the constriction than





During the time  $\Delta t$ , water entering the small section of pipe travels a distance  $v_1 \Delta t$ , while water leaving the large section goes a distance  $v_2 \Delta t$ . Since the same amount of water must enter as leave, the entrance volume  $A_1 \Delta x_1$  must equal the exit volume  $A_2 \Delta x_2$ . This gives  $A_1 v_1 \Delta t = A_2 v_2 \Delta t$ , or the result  $A_1 v_1 = A_2 v_2$  which is one form of the continuity equation.

in the wide sections. This way, the same volume of water per second gets past the constriction as passes per second past a wide section of the channel. Applying this idea to Figure (4), we see why the velocity vectors are longer, the fluid speed higher, in the narrow sections of the streamline channels than in the wide sections.

It is not too hard to go from the qualitative idea that fluid must flow faster in the narrow sections of a channel, to a quantitative result that allows us to calculate how much faster. In Figure (5), we are considering a section of streamline or flow tube which has an entrance area  $A_1$ , and exit area  $A_2$  as shown. In a short time  $\Delta t$ , the fluid at the entrance travels a distance  $\Delta x_1 = v_1 \Delta t$ , while at the exit the fluid goes a distance  $\Delta x_2 = v_2 \Delta t$ .

The volume of water that entered the stream during the time  $\Delta t$  is the shaded volume at the left side of the diagram, and is equal to the area  $A_1$  times the distance  $\Delta x_1$  that the fluid has moved

Volume of water  
entering in 
$$\Delta t$$
 =  $A_1 \Delta x_1 = A_1 v_1 \Delta t$  (1)

The volume of water leaving the same amount of time is

Volume of water   
leaving during 
$$\Delta t$$
 =  $A_2 \Delta x_2 = A_2 v_2 \Delta t$  (2)

If the water does not get squeezed up or compressed inside the stream between  $A_1$  and  $A_2$ , if we have an *incompressible fluid*, which is quite true for water and in many cases even true for air, then the volume of fluid entering and the volume of the fluid leaving during the time  $\Delta t$  must be equal. Equating Equations (1) and (2) and cancelling the  $\Delta t$  gives

$$\mathbf{A}_1 \mathbf{v}_1 = \mathbf{A}_2 \mathbf{v}_2 \qquad continuity \ equation \qquad (3)$$

Equation (3) is known as the *continuity equation* for incompressible fluids. It is a statement that we do not squeeze up or lose any fluid in the stream. It also tells us that the velocity of the fluid is inversely proportional

to the cross sectional area of the stream at that point. If the cross sectional area in a constriction has been cut in half, then the speed of the water must double in order to get the fluid through the constriction.

If we have a map of the streamlines, and know the entrance speed  $v_1$  of the fluid, then we can determine the magnitude and direction of the fluid velocity  $\vec{v}_2$  at any point downstream. The direction of  $\vec{v}_2$  is parallel to the streamline at Point (2), and the magnitude is given by  $v_2 = v_1 (A_1/A_2)$ . Thus a careful map of the fluid streamlines, combined with the continuity equation, give us almost a complete picture of the fluid motion. The only additional information we need is the entrance speed.



#### Figure 6

Point source of water. Imagine that water molecules are created inside the small sphere and flow radially out through its surface at a speed  $v_1$ . The same molecules will eventually flow out through the larger sphere at a lesser speed  $v_2$ . If no water molecules are created or destroyed outside the small sphere, then the continuity equation  $A_1v_1 = A_2v_2$  requires that  $(4\pi r_1^2)v_1 = (4\pi r_2^2)v_2$ .

## **Velocity Field of a Point Source**

This is an artificial example that shows us how to apply the continuity equation in a somewhat unexpected way, and leads to some ideas that will be very important in our later discussion of electric fields.

For this example, imagine a small magic sphere that creates water molecules inside and lets the water molecules flow out through the surface of the sphere. (Or there may be an unseen hose that supplies the water that flows out through the surface of the sphere.)

Let the small sphere have a radius  $r_1$ , area  $4\pi r_1^2$  and assume that the water is emerging radially out through the small sphere at a speed  $v_1$  as shown in Figure (6). Also let us picture that the small sphere is at the center of a huge swimming pool full of water, that the sides of the pool are so far away that the water continues to flow radially outward at least for several meters. Now conceptually construct a second sphere of radius  $r_2 > r_1$  centered on the small sphere as in Figure (6).

During one second, the volume of water flowing out of the small sphere is  $v_1A_1$ , corresponding to  $\Delta t = 1$  sec in Equation (1). By the continuity equation, the volume of water flowing out through the second sphere in one second,  $v_2A_2$ , must be the same in order that no water piles up between the spheres. Using the fact that  $A_1 = 4\pi r_1^2$  and  $A_2 = 4\pi r_2^2$ , we get

$$v_1 A_1 = v_2 A_2 \qquad \begin{array}{c} continuity\\ equation \end{array}$$
$$v_1 4\pi r_1^2 = v_2 4\pi r_2^2$$
$$v_2 = \frac{1}{r_2^2} (v_1 r_1^2) \qquad (4a)$$

Equation 4a tells us that as we go out from the "magic sphere", as the distance  $r_2$  increases, the velocity  $v_2$  drops off as the inverse square of  $r_2$ , as  $1/r_2^2$ . We can write this relationship in the form

$$v_2 \propto \frac{1}{r_2^2}$$
 The symbol  
 $\propto means$  "proportional to" (4b)

A small spherical source like that shown in Figure (6) is often called a *point source*. We see that a point source of water produces a  $1/r^2$  velocity field.

## Velocity Field of a Line Source

One more example which we will often use later is the *line source*. This is much easier to construct than the point source where we had to create water molecules. Good models for a line source of water are the sprinkling hoses used to water gardens. These hoses have a series of small holes that let the water flow radially outward.

For this example, imagine that we have a long sprinkler hose running down the center of an immense swimming pool. In Figure (7) we are looking at a cross section of the hose and see a radial flow that looks very much like Figure (6). The side view, however, is different. Here we see that we are dealing with a line rather than a point source of water.

Consider a section of the hose and fluid of length L. The volume of water flowing in one second out through this section of hose is  $v_1A_1$  where  $A_1 = L$  times (the circumference of the hose) =  $L(2\pi r_1)$ .

Volume of  
water / sec  
from a section  
L of hose 
$$\left. \begin{array}{l} = v_1 A_1 = v_1 L 2\pi r_1 \\ \end{array} \right.$$

If the swimming pool is big enough so that this water continues to flow radially out through a cylindrical area  $A_2$  concentric with and surrounding the hose, then the volume of water per second (we will call this the "flux" of water) out through  $A_2$  is

Volume of  
water / sec out  
through 
$$A_2$$
 =  $v_2A_2 = v_2(2\pi r_2L)$ 

Using the continuity equation to equate these volumes of water per second gives

$$v_{1}A_{1} = v_{2}A_{2}$$

$$v_{1}L2\pi r_{1} = v_{2}L2\pi r_{2}; \quad v_{1}r_{1} = v_{2}r_{2}$$

$$v_{2} = \frac{(v_{1}r_{1})}{r_{2}} \propto \frac{1}{r_{2}}$$
(5)

We see that the velocity field of a line source drops off as 1/r rather than  $1/r^2$  which we got from a point source.



a) End view of line source



b) Side view of line source

#### Figure 7

Line source of water. In a line source, the water flows radially outward through a cylindrical area whose length we choose as L and whose circumference is  $2\pi r$ .

## FLUX

Sometimes simply changing the name of a quantity leads us to new ways of thinking about it. In this case we are going to use the word *flux* to describe the *amount of water flowing per second* out of some volume. From the examples we have considered, the flux of water out through volumes  $V_1$  and  $V_2$  are given by the formulas

$$\begin{array}{c} \text{Flux of} \\ \text{water} \\ \text{out of } V_1 \end{array} \end{array} \equiv \left\{ \begin{array}{c} \text{Volume of} \\ \text{water flowing} \\ \text{per second} \\ \text{out of } V_1 \end{array} \right\} = v_1 A_1 \\ \end{array} \\ \begin{array}{c} \text{Flux of} \\ \text{water} \\ \text{out of } V_2 \end{array} \right\} = v_2 A_2 \\ \end{array}$$

The continuity equation can be restated by saying that the flux of water out of  $V_1$  must equal the flux out of  $V_2$ if the water does not get lost or compressed as it flows from the inner to the outer surface.

So far we have chosen simple surfaces, a sphere and a cylinder, and for these surfaces the flux of water is simply the fluid speed v times the area out through which it is flowing. Note that for our cylindrical surface shown in Figure (8), no water is flowing out through the ends of the cylinder, thus only the outside area  $(2\pi rL)$  counted in our calculation of flux. A more general way of stating how we calculate flux is to say that it is the fluid speed v times the *perpendicular area* A<sub>⊥</sub> through which the fluid is flowing. For the cylinder, the perpendicular area A<sub>⊥</sub> is the outside area  $(2\pi rL)$ ; the ends of the cylinder are parallel to the flow and therefore do not count.



#### Figure 8

With a line source, all the water flows through the cylindrical surface surrounding the source and none through the ends. Thus  $A_{\perp}$ , the perpendicular area through which the water flows is  $(2\pi r) \times L$ .

The concept of flux can be generalized to irregular flows and irregularly shaped surfaces. To handle that case, break the flow up into a bunch of small flow tubes separated by streamlines, construct a perpendicular area for each flow tube as shown in Figure (9), and then calculate the total flux by adding up the fluxes from each flow tube.

Total Flux = 
$$v_1A_1 + v_2A_2 + v_2A_2 + \dots$$
  
=  $\sum_i v_i A_{\perp}^i$  (6)

In the really messy cases, the sum over flow tubes becomes an integral as we take the limit of a large number of infinitesimal flow tubes.

For this text, we have gone too far. We will not work with very complicated flows. We can learn all we want from the simple ones like the flow out of a sphere or a cylinder. In those cases the perpendicular area is obvious and the flux easy to calculate. For the spherical flow of Figure (6), we see that the velocity field dropped off as  $1/r^2$  as we went out from the center of the sphere. For the cylinder in Figure (7) the velocity field dropped off less rapidly, as 1/r.



#### Figure 9

To calculate the flux of water in an arbitrarily shaped flow break up the flow into many small flux tubes where the fluid velocity is essentially uniform across the small tube as shown. The flux through the i-th tube is simply  $v_iA_i$ , and the total flux is the sum of the fluxes  $\Sigma_i v_iA_i$  through each tube.

►⊽

## **BERNOULLI'S EQUATION**

Our discussion of flux was fairly lengthy, not so much for the results we got, but to establish concepts that we will use extensively later on in our discussion of electric fields. Another topic, Bernoulli's law, has a much more direct application to the understanding of fluid flows. It also has some rather surprising consequences which help explain why airplanes can fly and how a sailboat can sail up into the wind.

Bernoulli's law involves an energy relationship between the pressure, the height, and the velocity of a fluid. The theorem assumes that we have a constant density fluid moving with a steady flow, and that viscous effects are negligible, as they often are for fluids such as air and water. Consider a small tube of flow bounded by streamlines as shown in Figure (10). In a short time  $\Delta t$  a small volume of fluid enters on the left and an equal volume exits on the right. If the exiting volume has more energy than the entering volume, the extra energy had to come from the work done by pressure forces acting on the fluid in the flow tube. Equating the work done by the pressure forces to the increase in energy gives us Bernoulli's equation.

To help visualize the situation, imagine that the streamline boundaries of the flow tube are replaced by frictionless, rigid walls. This would have no effect on the flow of the fluid, but focuses our attention on the ends of the tube where the fluid is flowing in on the left, at what we will call Point (1), and out on the right at Point (2).



#### Figure 10

Derivation of Bernoulli's equation. Select a flow tube bounded by streamlines. For the steady flow of an incompressible fluid, during a time  $\Delta t$  the same volume of fluid must enter on the left as leave on the right. If the exiting fluid has more energy than the entering fluid, the increase must be a result of the net work done by the pressure forces acting on the fluid.

As a further aid to visualization, imagine that a small frictionless cylinder is temporarily inserted into the entrance of the tube as shown in Figure (11a), and at the exit as shown in Figure (11b). Such cylinders have no effect on the flow but help us picture the pressure forces.

At the entrance, if the fluid pressure is  $P_1$  and the area of the cylinder is  $A_1$ , then the external fluid exerts a net force of magnitude

$$\left|\vec{\mathbf{F}}_{1}\right| = \mathbf{P}_{1}\mathbf{A}_{1} \tag{7a}$$

directed perpendicular to the surface of the cylinder as shown. We can think of this force  $\vec{F}_1$  as the pressure force that the outside fluid exerts on the fluid inside the flow tube. At the exit, the external fluid exerts a pressure force  $\vec{F}_2$  of magnitude

$$\left|\vec{\mathbf{F}}_{2}\right| = \mathbf{P}_{2}\mathbf{A}_{2} \tag{7b}$$

directed perpendicular to the piston, i.e., back toward the fluid inside the tube.



Figure 11

The flow would be unchanged if we temporarily inserted frictionless pistons at the entrance and exit.

Thus the fluid inside the flow tube is subject to external pressure forces,  $\vec{F}_1$  in from the left and  $\vec{F}_2$  in from the right. During a time  $\Delta t$ , the fluid at the entrance moves a distance  $\Delta x_1 = v_1 \Delta t$  as shown in Figure (10). While moving this distance, the entering fluid is subject to the pressure force  $\vec{F}_1$ , thus the work  $\Delta W_1$  done by the pressure force at the entrance is

$$\Delta \mathbf{W}_{1} = \vec{\mathbf{F}}_{1} \cdot \Delta \vec{\mathbf{x}}_{1} = \left( \mathbf{P}_{1} \mathbf{A}_{1} \right) \left( \mathbf{v}_{1} \Delta t \right)$$
(8a)

At the exit, the fluid moves out a distance  $\Delta x_2 = v_2 \Delta t$ , while the external force pushes back in with a pressure force  $\vec{F}_2$ . Thus the pressure forces do *negative* work on the inside fluid, with the result

$$\Delta W_2 = \vec{F}_2 \cdot \Delta \vec{x}_2 = -(P_2 A_2)(v_2 \Delta t)$$
(8b)

The net work  $\Delta W$  done during a time  $\Delta t$  by external pressure forces on fluid inside the flow tube is therefore

$$\Delta W = \Delta W_1 + \Delta W_2$$
  
=  $P_1 (A_1 v_1 \Delta t) - P_2 (A_2 v_2 \Delta t)$  (9)

Equation (9) can be simplified by noting that  $A_1v_1\Delta t = A_1\Delta x_1$  is the volume  $\Delta V_1$  of the entering fluid. Likewise  $A_2v_2\Delta t = A_2\Delta x_2$  is the volume  $\Delta V_2$  of the exiting fluid. But during  $\Delta t$ , the same volume  $\Delta V$  of fluid enters and leaves, thus  $\Delta V_1 = \Delta V_2 = \Delta V$  and we can write Equation (9) as

work done  
by external  
pressure  
forces  
on fluid  
inside flow  
tube 
$$\Delta W = \Delta V (P_1 - P_2)$$
 (10)

The next step is to calculate the change in energy of the entering and exiting volumes of fluid. The energy  $\Delta E_1$  of the entering fluid is its kinetic energy  $1/2(\Delta m)v_1^2$  plus its gravitational potential energy  $(\Delta m)gh_1$ , where  $\Delta m$  is the mass of the entering fluid. If the fluid has a density  $\rho$ , then  $\Delta m = \rho \Delta V$  and we get

(15)

$$\Delta E_{1} = \frac{1}{2} (\rho \Delta V) v_{1}^{2} + (\rho \Delta V) gh_{1}$$

$$= \Delta V \left( \frac{1}{2} \rho v_{1}^{2} + \rho gh_{1} \right)$$
(11a)

At the exit, the same mass and volume of fluid leave in time  $\Delta t$ , and the energy of the exiting fluid is

$$\Delta E_2 = \Delta V \left( \frac{1}{2} \rho v_2^2 + \rho g h_2 \right)$$
(11b)

The change  $\Delta E$  in the energy in going from the entrance to the exit is therefore

$$\Delta \mathbf{E} = \Delta \mathbf{E}_2 - \Delta \mathbf{E}_1$$
$$= \Delta \mathbf{V} \left( \frac{1}{2} \rho \mathbf{v}_2^2 + \rho g \mathbf{h}_2 - \frac{1}{2} \rho \mathbf{v}_1^2 - \rho g \mathbf{h}_1 \right) \quad (12)$$

Equating the work done, Equation (10) to the change in energy, Equation (12) gives

$$\Delta V(P_1 - P_2) = \Delta V \left(\frac{1}{2}\rho v_2^2 + \rho g h_2 - \frac{1}{2}\rho v_1^2 - \rho g h_1\right)$$
(13)

Not only can we cancel the  $\Delta Vs$  in Equation (13), but we can rearrange the terms to make the result easier to remember. We get

$$P_1 + \frac{1}{2}\rho v_1^2 + \rho g h_1 = P_2 + \frac{1}{2}\rho v_2^2 + \rho g h_2$$
 (14)

In this form, an interpretation of Bernoulli's equation begins to emerge. We see that the quantity  $(P + \rho gh + \frac{1}{2}\rho v^2)$  has the same numerical value at the entrance, Point (1), as at the exit, Point (2). Since we can move the starting and ending points anywhere along the flow tube, we have the more general result

$$\left(P + \rho g h + \frac{1}{2}\rho v^2\right) = \begin{cases} \text{constant anywhere} \\ \text{along a flow tube} \\ \text{or streamline} \end{cases}$$

Equation (15) is our final statement of Bernoulli's equation. In words it says that for the steady flow of an incompressible, non viscous, fluid, the quantity  $(P + \rho gh + 1/2\rho v^2)$  has a constant value along a streamline.

The restriction that  $(P + \rho gh + 1/2\rho v^2)$  is constant along a streamline has to be taken seriously. Our derivation applied energy conservation to a plug moving along a small flow tube whose boundaries are streamlines. We did not consider plugs of fluid moving in different flow tubes, i.e., along different streamlines. special flows, the quantity For some  $P + \rho gh + 1/2\rho v^2$  has the same value throughout the entire fluid. But for most flows,  $P + \rho gh + 1/2\rho v^2$  has different values on different streamlines. Since we haven't told you what the special flows are, play it safe and assume that the numerical value of  $(P + \rho gh + 1/2\rho v^2)$  can change when you hop from one streamline to another.

## APPLICATIONS OF BERNOULLI'S EQUATION

Bernoulli's equation is a rather remarkable result that some quantity  $(P + \rho gh + 1/2\rho v^2)$  has a value that doesn't change as you go along a streamline. The terms inside, except for the P term, look like the energy of a unit volume of fluid. The P term came from the work part of the energy conservation theorem, and cannot strictly be interpreted as some kind of pressure energy. As tempting as it is to try to give an interpretation to the terms in Bernoulli's equation, we will put that off for a while until we have worked out some practical applications of the formula. Once you see how much the equation can do, you will have a greater incentive to develop an interpretation.

## **Hydrostatics**

Let us start with the simplest application of Bernoulli's equation, namely the case where the fluid is at rest. In a sense, all the fluid is on the same streamline, and we have

$$P + \rho gh = \begin{cases} constant \\ throughout \\ the fluid \end{cases}$$
(16)

Suppose we have a tank of water shown in Figure (12). Let the pressure be atmospheric pressure at the surface, and set h = 0 at the surface. Therefore at the surface



Figure 12 Hydrostatic pressure at a depth y is atmospheric pressure plus  $\rho gy$ .

 $P_{at} + \rho g(0) = constant$ 

and the constant is  $P_{at}$ . For any depth y = -h, we have

$$P - \rho gy = constant = P_{at}$$

$$P = P_{at} + \rho gy$$
(17)

We see that the increase in pressure at a depth y is  $\rho gy$ , a well-known result from hydrostatics.

## Exercise 1

The density of water is  $\rho = 10^3 \text{Kg/m}^3$  and atmospheric pressure is  $P_{at} = 1.0 \times 10^5 \text{N/m}^2$ . At what depth does a scuba diver breath air at a pressure of 2 atmospheres? (At what depth does  $\rho gy = P_{at}$ ?) (Your answer should be 10.2m or 33 ft.)

## Exercise 2

What is the pressure, in atmospheres, at the deepest part of the ocean? (At a depth of 8 kilometers.)

## Leaky Tank

For a slightly more challenging example, suppose we have a tank filled with water as shown in Figure (13). A distance h below the surface of the tank we drill a hole and the water runs out of the hole at a speed v. Use Bernoulli's equation to determine the speed v of the exiting water.



#### Figure 13

Water squirting out through a hole in a leaky tank. A streamline connects the leak at Point (2) with some Point (1) on the surface. Bernoulli's equation tells us that the water squirts out at the same speed it would have if it had fallen a height h.

**Solution:** Somewhere there will be a streamline connecting the free surface of the water (1) to a Point (2) in the exiting stream. Applying Bernoulli's equation to Points (1) and (2) gives

$$P_1 + \rho g h_1 + \frac{1}{2} \rho v_1^2 = P_2 + \rho g h_2 + \frac{1}{2} \rho v_2^2$$

Now  $P_1 = P_2 = P_{at}$ , so the Ps cancel. The water level in the tank is dropping very slowly, so that we can set  $v_1 = 0$ . Finally  $h_1 - h_2 = h$ , and we get

$$\frac{1}{2}\rho v_2^2 = \rho g (h_1 - h_2) = \rho g h$$
 (18)

The result is that the water coming out of the hole is moving just as fast as it would if it had fallen freely from the top surface to the hole we drilled.

## **Airplane Wing**

In the example of a leaky tank, Bernoulli's equation gives a reasonable, not too exciting result. You might have guessed the answer by saying energy should be conserved. Now we will consider some examples that are more surprising than intuitive. The first explains how an airplane can stay up in the air.

Figure (14) shows the cross section of a typical airplane wing and some streamlines for a typical flow of fluid around the wing. (We copied the streamlines from our demonstration in Figure 3).

The wing is purposely designed so that the fluid has to flow farther to get over the top of the wing than it does to flow across the bottom. To travel this greater distance, the fluid has to move faster on the top of the wing (at Point 1), than at the bottom (at Point 2).



#### Figure 14

Streamline flow around an airplane wing. The wing is shaped so that the fluid flows faster over the top of the wing, Point (1) than underneath, Point (2). As a result the pressure is higher beneath Point (2) than above Point (1). Arguing that the fluid at Point (1) on the top and Point (2) on the bottom started out on essentially the same streamline (Point 0), we can apply Bernoulli's equation to Points (1) and (2) with the result

$$P_1 + \frac{1}{2}\rho v_1^2 + \rho g h_1 = P_2 + \frac{1}{2}\rho v_2^2 + \rho g h_2$$

We have crossed out the  $\rho$ gh terms because the difference in hydrostatic pressure  $\rho$ gh across the wing is negligible for a light fluid like air.

Here is the important observation. Since the fluid speed  $v_1$  at the top of the wing is higher than the speed  $v_2$  at the bottom, *the pressure*  $P_2$  *at the bottom must be greater than*  $P_1$  *at the top* in order that the sum of the two terms  $(P + 1/2\rho v^2)$  be the same. The extra pressure on the bottom of the wing is what provides the lift that keeps the airplane up in the air.

There are two obvious criticisms of the above explanation of how airplanes get lift. What about stunt pilots who fly upside down? And how do balsa wood gliders with flat wings fly? The answer lies in the fact that the shape of the wing cross-section is only one of several important factors determining the flow pattern around a wing.

Figure (15) is a sketch of the flow pattern around a flat wing flying with a small angle of attack  $\theta$ . By having an angle of attack, the wing creates a flow pattern where the streamlines around the top of the wing are longer than those under the bottom. The result is that the fluid flows faster over the top, therefore the pressure must be lower at the top (higher at the bottom) and we still get lift. The stunt pilot flying upside down must fly with a great enough angle of attack to overcome any downward lift designed into the wing.



#### Figure 15

A balsa wood model plane gets lift by having the wing move forward with an upward tilt, or angle of attack. The flow pattern around the tilted wing gives rise to a faster flow and therefore reduced pressure over the top.

## Sailboats

Sailboats rely on Bernoulli's principle not only to supply the "lift" force that allows the boat to sail into the wind, but also to create the "wing" itself. Figure (16) is a sketch of a sailboat heading at an angle  $\theta$  off from the wind. If the sail has the shape shown, it looks like the airplane wing of Figure (14), the air will be moving faster over the outside curve of the sail (Position 1) than the inside (Position 2), and we get a higher pressure on the inside of the sail. This higher pressure on the inside both pushes the sail cloth out to give the sail an airplane wing shape, and creates the lift force shown in the diagram. This lift force has two components. One pulls the boat forward. The other component, however, tends to drag the boat sideways. To prevent the boat from slipping sideways, sailboats are equipped with a centerboard or a keel.

The operation of a sailboat is easily demonstrated using an air cart, glider and fan. Mount a small sail on top of the air cart glider (the light plastic shopping bags make excellent sail material) and elevate one end of the cart as shown in Figure (17) so that the cart rests at the low end. Then mount a fan so that the wind blows down and across as shown. With a little adjustment of the angle of the fan and the tilt of the air cart, you can observe the cart sail up the track, into the wind.

If you get the opportunity to sail a boat, remember that it is the Bernoulli effect that both shapes the sail and propels the boat. Try to adjust the sail so that it has a good airplane wing shape, and remember that the higher speed wind on the outside of the sail creates a low pressure that sucks the sailboat forward. You'll go faster if you keep these principles in mind.



#### Figure 16

A properly designed sail takes on the shape of an airplane wing with the wind traveling faster, creating a lower pressure on the outside of the sail (Point 1). This low pressure on the outside both sucks the canvas out to maintain the shape of the sail and provides the lift force. The forward component of the lift force moves the boat forward and the sideways component is offset by the water acting on the keel.



#### Figure 17

Sailboat demonstration. It is easy to rig a mast on an air cart, and use a small piece of a light plastic bag for a sail. Place the cart on a tilted air track so that the cart will naturally fall backward. Then turn on a fan as shown, and the cart sails up the track into the wind.

## The Venturi Meter

Another example, often advertised as a simple application of Bernoulli's equation, is the Venturi meter shown in Figure (18). We have a tube with a constriction, so that its cross-sectional area  $A_1$  at the entrance and the exit, is reduced to  $A_2$  at the constriction. By the continuity equation (3), we have

$$v_1A_1 = v_2A_2; \quad v_2 = \frac{v_1A_1}{A_2}$$

As expected, the fluid travels faster through the constriction since  $A_1 > A_2$ .

Now apply Bernoulli's equation to Points (1) and (2). Since these points are at the same height, the  $\rho$ gh terms cancel and Bernoulli's equation becomes

$$P_1 + \frac{1}{2}\rho v_1^2 = P_2 + \frac{1}{2}\rho v_2^2$$

Since  $v_2 > v_1$ , the pressure  $P_2$  in the constriction must be less than the pressure  $P_1$  in the main part of the tube. Using  $v_2 = v_1A_1/A_2$ , we get

$$\begin{cases} \text{Pressure} \\ \text{drop in} \\ \text{constriction} \end{cases} = \left( P_1 - P_2 \right) \\ = \frac{1}{2} \rho \left( v_2^2 - v_1^2 \right) \\ = \frac{1}{2} \rho \left( v_1^2 A_1^2 / A_2^2 - v_1^2 \right) \\ = \frac{1}{2} \rho v_1^2 \left( A_1^2 / A_2^2 - 1 \right) \end{cases}$$
(19)

To observe the pressure drop, we can mount small tubes (A) and (B) as shown in Figure (18), to act as barometers. The lower pressure in the constriction will



Figure 18

Venturi meter. Since the water flows faster through the constriction, the pressure is lower there. By using vertical tubes to measure the pressure drop, and using Bernoulli's equation and the continuity equation, you can determine the flow speeds  $v_1$  and  $v_2$ . cause the fluid level in barometer (B) to be lower than in the barometer over the slowly moving, high pressure stream. The height difference h means that there is a pressure difference

$$\frac{\text{Pressure}}{\text{difference}} = (P_1 - P_2) = \rho gh$$
(20)

If we combine Equations (19) and (20),  $\rho$  cancels and we can solve for the speed v<sub>1</sub> of the fluid in the tube in terms of the quantities g, h, A<sub>1</sub> and A<sub>2</sub>. The result is

$$v_1 = \sqrt{\frac{2gh}{\left(A_1^2/A_2^2 - 1\right)}}$$
 (21)

Because we can determine the speed  $v_1$  of the main flow by measuring the height difference h of the two columns of fluid, the setup in Figure (18) forms the basis of an often used meter to measure fluid flows. A meter based on this principle is called a *Venturi* meter.

## **Exercise 3**

Show that all the terms in Bernoulli's equation have the same dimensions. (Use MKS units.)

## **Exercise 4**

In a classroom demonstration of a venturi meter shown in Figure (18a), the inlet and outlet pipes had diameters of 2 cm and the constriction a diameter of 1 cm. For a certain flow, we noted that the height difference h in the barometer tubes was 7 cm. How fast, in meters/sec, was the fluid flowing in the inlet pipe?



#### Figure 18a

Venturi demonstration. We see about a 7cm drop in the height of the barometer tubes at the constriction.

## **The Aspirator**

In Figure (18), the faster we move the fluid through the constriction (the greater  $v_1$  and therefore  $v_2$ ), the greater the height difference h in the two barometer columns. If we turn  $v_1$  up high enough, the fluid is moving so fast through section 2 that the pressure becomes negative and we get suction in barometer 2. For even higher speed flows, the suction at the constriction becomes quite strong and we have effectively created a crude vacuum pump called an *aspirator*. Typically aspirators like that shown in Figure (19) are mounted on cold water faucets in chemistry labs and are used for sucking up various kinds of fluids.



Figure 19a

If the water flows through the constriction fast enough, you get a negative pressure and suction in the attached tube.



Figure 19b

If the constriction is placed on the end of a water faucet as shown, you have a device called an aspirator that is often used in chemistry labs for sucking up fluids.

## Care in Applying Bernoulli's Equation

Although the Venturi meter and aspirator are often used as simple examples of Bernoulli's equation, considerable care must be used in applying Bernoulli's equation in these examples. To illustrate the trouble you can get into, suppose you tried to apply Bernoulli's equation to Points (3) and (4) of Figure (20). You would write

$$P_{3} + \rho g h_{3} + \frac{1}{2} \rho v_{3}^{2} = P_{4} + \rho g h_{4} + \frac{1}{2} \rho v_{4}^{2}$$
(22)

Now  $P_3 = P_4 = P_{atmosphere}$  because Points (3) and (4) are at the liquid surface. In addition the fluid is at rest in tubes (3) and (4), therefore  $v_3 = v_4 = 0$ . Therefore Bernoulli's equation predicts that

$$\rho gh_3 = \rho gh_4$$

or that  $h_3 = h_4$  and there should be no height difference!

What went wrong? The mistake results from the fact that no streamlines go from position (3) to position (4), and therefore Bernoulli's equation does not have to apply. As shown in Figure (21) the streamlines flow across the bottom of the barometer tubes but do not go up into them. It turns out that we cannot apply Bernoulli's equation across this break in the streamlines. It requires some experience or a more advanced knowledge of hydrodynamic theory to know that you can treat the little tubes as barometers and get the



#### Figure 20

If you try to apply Bernoulli's equation to Points (3) and (4), you predict, incorrectly, that Points (3) and (4) should be at the same height. The error is that Points (3) and (4) do not lie on the same streamline, and therefore you cannot apply Bernoulli's equation to them.



#### Figure 21

The water flows past the bottom of the barometer tube, not up into the tube. Thus Point (3) is not connected to any of the streamlines in the flow. The vertical tube acts essentially as a barometer, measuring the pressure of the fluid flowing beneath it.

correct answer. Most texts ignore this complication, but there are always some students who are clever enough to try to apply Bernoulli's equation across the break in the flow at the bottom of the small tubes and then wonder why they do not get reasonable answers.

There is a remarkable fluid called *superfluid helium* which under certain circumstances will not have a break in the flow at the base of the barometer tubes. (Superfluid helium is liquefied helium gas cooled to a temperature below  $2.17 \,^{\circ}$  K). As shown in Figure (22) the streamlines actually go up into the barometer tubes, Points (3) and (4) are connected by a streamline, Bernoulli's equation should apply and we should get no height difference. This experiment was performed in 1965 by Robert Meservey and the heights in the two barometer tubes were just the same!



#### Figure 22

In superfluid helium, the streamlines actually go up into the barometer tubes and Bernoulli's equation can be applied to Points (3) and (4). The result is that the heights of the fluid are the same as predicted. (Experiment by R. Meservey, see Physics Of Fluids, July 1965.)

#### Hydrodynamic Voltage

When we studied the motion of a projectile, we found that the quantity  $(1/2 \text{ mv}^2 + \text{mgh})$  did not change as the ball moved along its parabolic trajectory. When physicists discover a quantity like  $(1/2 \text{ mv}^2 + \text{mgh})$  that does not change, they give that quantity a name, in this case "*the ball's total energy*", and then say that they have discovered a new law, namely "the ball's total energy is conserved as the ball moves along its trajectory".

With Bernoulli's equation we have a quantity  $(P + \rho gh + 1/2\rho v^2)$  which is constant along a streamline when we have the steady flow of an incompressible, non viscous fluid. Here we have a quantity  $(P + \rho gh + 1/2\rho v^2)$  that is conserved under special circumstances; perhaps we should give this quantity a name also.

The term  $\rho$ gh is the gravitational potential energy of a unit volume of the fluid, and  $1/2\rho v^2$  is the same volume's kinetic energy. Thus our Bernoulli term has the dimensions and characteristics of the energy of a unit volume of fluid. But the pressure term, which came from the work part of the derivation of Bernoulli's equation, is not a real energy term. There is no pressure energy P stored in an incompressible fluid, and Bernoulli's equation is not truly a statement of energy conservation for a unit volume of fluid.

However, as we have seen, the Bernoulli term is a useful concept, and deserves a name. Once we name it, we can say that "" is conserved along a streamline under the right circumstances. Surprisingly there is not an extensive tradition for giving the Bernoulli term a name so that we have to concoct a name here. At this point our choice of name will seem a bit peculiar, but it is chosen with later discussions in mind. We will call the Bernoulli term *hydrodynamic voltage* 

$$\begin{vmatrix} \text{Hydrodynamic} \\ \text{Voltage} \end{vmatrix} \equiv \left( P + \rho g h + \frac{1}{2} \rho v^2 \right) \qquad (23)$$

and Bernoulli's equation states that *the hydrodynamic* voltage of an incompressible, non viscous fluid is constant along a streamline when the flow is steady.

We obviously did not invent the word voltage; the name is commonly used in discussing electrical devices like high voltage wires and low voltage batteries. It turns out that there is a precise analogy between the concept of voltage used in electricity theory, and the Bernoulli term we have been discussing. To emphasize the analogy, we are naming the Bernoulli term hydrodynamic voltage. The word "hydrodynamic" is included to remind us that we are missing some of the electrical terms in a more general definition of voltage. We are discussing hydrodynamic voltage before electrical voltage because hydrodynamic voltage involves fluid concepts that are more familiar, easier to visualize and study, than the corresponding electrical concepts.

## **Town Water Supply**

One of the familiar sights in towns where there are no nearby hills is the water tank somewhat crudely illustrated in Figure (23). Water is pumped from the reservoir into the tank to fill the tank up to a height h as shown.

For now let us assume that all the pipes attached to the tanks are relatively large and frictionless so that we can neglect viscous effects and apply Bernoulli's equation to the water at the various points along the water system. At Point (1), the pressure is simply atmospheric pressure  $P_{at}$ , the water is essentially not flowing, and the hydrodynamic voltage consists mainly of  $P_{at}$  plus the gravitational term gh<sub>1</sub>

$$\begin{array}{l} \text{Hydrodynamic} \\ \text{Voltage} \end{array} \right\}_{1} = P_{\text{at}} + gh_{1}$$

By placing the tank high up in the air, the  $gh_1$  term can be made quite large. We can say that the tank gives us "high voltage" water.



Figure 23

The pressure in the town water supply may be maintained by pumping water into a water tank as shown. If the pipes are big enough we can neglect the viscous effect and apply Bernoulli's equation throughout the system, including the break in the water pipe at Point (3), and the top of the fountain, Point (4). Bernoulli's equation tells us that the hydrodynamic voltage of the water is the same at all the points along the water system. The purpose of the water tank is to ensure that we have high voltage water throughout the town. For example, at Point (2) at one of the closed faucets in the second house, there is no height left  $(h_2 = 0)$  and the water is not flowing. Thus all the voltage shows up as high pressure at the faucet.

$$\begin{array}{l} \text{Hydrodynamic} \\ \text{Voltage} \end{array} \right\}_2 = P_2$$

At Point (3) we have a break in the pipe and water is squirting up. Just above the break the pressure has dropped to atmospheric pressure and there is still no height. At this point the voltage appears mainly in the form of kinetic energy.

$$\begin{cases} \text{Hydrodynamic} \\ \text{Voltage} \end{cases}_3 = P_{at} + \frac{1}{2}\rho v^2 \end{cases}$$

Finally at Point (4) the water from the break reaches its maximum height and comes to rest before falling down again. Here it has no kinetic energy, the pressure is still atmospheric, and the hydrodynamic voltage is back in the form of gravitational potential energy. If no voltage has been lost, if Bernoulli's equation still holds, then the water at Point (4) must rise to the same height as the water at the surface in the town water tank.

In some sense, the town water tank serves as a huge "battery" to supply the hydrodynamic voltage for the town water system.

## **Viscous Effects**

We said that the hydrodynamic analogy of voltage involves familiar concepts. Sometimes the concepts are too familiar. Has your shower suddenly turned cold when someone in the kitchen drew hot water for washing dishes; or turned hot when the toilet was flushed? Or been reduced to a trickle when the laundry was being washed? In all of these cases there was a pressure drop at the shower head of either the hot water, the cold water, or both. A pressure drop means that you are getting lower voltage water at the shower head than was supplied by the town water tank (or by your home pressure tank).

The hydrodynamic voltage drop results from the fact that you are trying to draw too much water through small pipes, viscous forces become important, and Bernoulli's equation no longer applies. Viscous forces always cause a drop in the hydrodynamic voltage. This voltage drop can be seen in a classroom demonstration, Figure (24), where we have inserted a series of small barometer tubes in a relatively small flow tube. If we run a relatively high speed stream of water through the flow tube, viscous effects become observable and the pressure drops as the water flows down the tube. The pressure drop is made clear by the decreasing heights of the water in the barometer tubes as we go downstream.



**Figure 18a repeated** *Venturi demonstration.* 



## Figure 24

If we have a fairly fast flow in a fairly small tube, viscosity causes a pressure drop, or as we are calling it, a "hydrodynamic voltage" drop down the tube. This voltage drop is seen in the decreasing heights of the water in the barometer tubes. (In our Venturi demonstration of Figure (18a), the heights are lower on the exit side than the entrance side due to viscosity acting in the constriction.)

## VORTICES

The flows we have been considering, water in a pipe, air past a sailboat sail, are tame compared to a striking phenomena seen naturally in the form of hurricanes and tornados. These are examples of a fluid motion called a *vortex*. They are an extension, to an atmospheric scale, of the common bathtub vortex like the one we created in the funnel seen in Figure (25).

Vortices have a fairly well-defined structure which is seen most dramatically in the case of the tornado (see Figures 29 and 30). At the center of the vortex is the core. The core of a bathtub vortex is the hollow tube of air that goes down the drain. In a tornado or water spout, the core is the rapidly rotating air. For a hurricane it is the eye, seen in Figures (27) and (28), which can be amazingly calm and serene considering the vicious winds and rain just outside the eye.



**Figure 25** Bathtub vortex in a funnel. We stirred the water before letting it drain out.



#### Figure 26

Vortices tend to have a circular velocity field about the core, a velocity field  $v_{\theta}$  whose strength tends to drop off as 1/r as you go out from the core.

Outside the core, the fluid goes around in a circular pattern, the speed decreasing as the distance from the center increases. It turns out that viscous effects are minimized if the fluid speed drops off as 1/r where r is the radial distance from the center of the core as shown in Figure (26). At some distance from the center, the speed drops to below the speed of other local disturbances and we no longer see the organized motion.

The tendency of a fluid to try to maintain a 1/r velocity field explains why vortices have to have a core. You cannot maintain a 1/r velocity field down to r = 0, for then you would have infinite velocities at the center. To avoid this problem, the vortex either throws the fluid out of the core, as in the case of the hollow bathtub vortex, or has the fluid in the core move as a solid rotating object ( $v_{\theta} = r \omega$ ) in the case of a tornado, or has a calm fluid when the core is large (i.e., viscous effects of the land are important) as in the case of a hurricane.

While the tornado is a very well organized example of a vortex, it has been difficult to do precise measurements of the wind speeds in a tornado. One of the best measurements verifying the 1/r velocity field was when a tornado hit a lumber yard, and a television station using a helicopter recorded the motion of sheets of 4' by 8' plywood that were scattered by the tornado. (Using doppler radar, a wind speed of 318 miles per hour was recorded in a tornado that struck Oklahoma city on May 3, 1999--a world wind speed record!)



**Figure 27** *Eye of hurricane Allen viewed from a satellite. (Photograph courtesy of A. F. Haasler.)* 





Figure 29 Tornado in Kansas.



**Figure 30** *A tornado over water is called a water spout.* 



Figure 28 Hurricane approaching the east coast of the U.S.

## **Quantized Vortices in Superfluids**

For precision, nothing beats the "quantized" vortex in superfluid helium. We have already mentioned that superfluid helium flows up and down the little barometer tubes in a Venturi meter, giving no height difference and nullifying the effectiveness of the device as a velocity meter. This happened because superfluid helium has <u>NO</u> viscosity (absolutely none as far as we can tell) and can therefore flow into tiny places where other fluids cannot move.

More surprising yet is the structure of a vortex in superfluid helium. The vortex has a core that is about one atomic diameter across (you can't get much smaller than that), and a precise 1/r velocity field outside the core. Even more peculiar is the fact that the velocity field outside the core is given by the formula

$$v_{\theta} = \frac{\kappa}{2\pi r}; \quad \kappa = \frac{h}{m_{He}}$$
 (24)

where  $\kappa$ , called the "circulation of the vortex", has the precisely known value  $h/m_{He}$ , where  $m_{He}$  is the mass of a helium atom, and h is an atomic constant known as **Planck's constant**. The remarkable point is that the strength of a helium vortex has a precise value determined by atomic scale constants. (This is why we say that vortices in superfluid helium are quantized.) When we get to the study of atoms, and particularly the Bohr theory of hydrogen, we can begin to explain why helium vortices have precisely the strength  $\kappa = h/m_{He}$ . For now, we are mentioning vortices in superfluid helium as examples of an ideal vortex with a well-defined core and a precise 1/r velocity field outside.

Quantized vortices of a more complicated structure also occur in superconductors and play an important role in the practical behavior of a superconducting material. The superconductors that carry the greatest currents, and are the most useful in practical applications, have quantized vortices that are pinned down and cannot move around. One of the problems in developing practical applications for the new high temperature superconductors is that the quantized vortices tend to move and cause energy losses. Pinning these vortices down is one of the main goals of current engineering research.

## Exercise 5

This was an experiment, performed in the 1970s to study how platelets form plaque in arteries. The idea was that platelets deposit out of the blood if the flow of blood is too slow. The purpose of the experiment was to design a flow where one could easily see where the plaque began to form and also know what the velocity of the flow was there.

The apparatus is shown in Figure (31). Blood flows down through a small tube and then through a hole in a circular plate that is suspended a small distance d above a glass plate. When the blood gets to the glass it flows radially outward as indicated in Figure (31c). As the blood flowed radially outward, its velocity decreases. At a certain radius, call it  $r_p$ , platelets began to deposit on the glass. The flow was photographed by a video camera looking up through the glass.

For this problem, assume that the tube radius was  $r_t = .4mm$ , and that the separation d between the circular plate and the glass was d = .5mm. If blood were flowing down the inlet tube at a rate of half a cubic centimeter per second, what is the average speed of the blood

a) inside the inlet tube?

b) at a radius  $r_p = 2cm$  out from the hole in the circular plate?

(By average speed, we mean neglect fluid friction at walls, and assume that the flow is uniform across the radius of the inlet pipe and across the gap as indicated in Figure (31d).

#### Exercise 6

A good review of both the continuity equation and Bernoulli's equation, is to derive on your own, without looking back at the text, the formula

$$v_1 = \sqrt{\frac{2gh}{A_1^2/A_2^2 - 1}}$$
(21)

for the flow speed in a venturi meter. The various quantities  $v_1$ , h,  $A_1$  and  $A_2$  are defined in Figure (18) reproduced on the opposite page. (If you have trouble with the derivation, review it in the text, and then a day or so later, try the derivation again on your own.



#### Figure 31 a,b,c

*Experiment to measure the blood flow velocity at which platelets stick to a glass plate. This is an application of the continuity equation.* 



#### Figure 31d

Neglect fluid friction at walls, and assume that the flow is uniform across the radius of the inlet pipe and across the gap



#### Figure 18

Venturi meter. Since the water flows faster through the constriction, the pressure is lower there. By using vertical tubes to measure the pressure drop, and using Bernoulli's equation and the continuity equation, you can determine the flow speeds  $v_1$  and  $v_2$ .

# Chapter 24 Coulomb's Law and Gauss' Law

In our discussion of the four basic interactions we saw that the electric and gravitational interaction had very similar  $1/r^2$  force laws, but produced very different kinds of structures. The gravitationally bound structures include planets, solar systems, star clusters, galaxies, and clusters of galaxies. Typical electrically bound structures are atoms, molecules, people, and redwood trees. Although the force laws are similar in form, the differences in the structures they create result from two important differences in the forces. Gravity is weaker, far weaker, than electricity. On an atomic scale gravity is so weak that its effects have not been seen. But electricity has both attractive and repulsive forces. On a large scale, the electric forces cancel so completely that the weak but non-cancelling gravity dominates astronomical structures.

#### COULOMB'S LAW

In Chapter 18 we briefly discussed Coulomb's electric force law, primarily to compare it with gravitational force law. We wrote Coulomb's law in the form

$$\left| \vec{F} \right| = \frac{KQ_1Q_2}{r^2}$$
 Coulomb,s Law (1)

where  $Q_1$  and  $Q_2$  are two charges separated by a distance r as shown in Figure (1). If the charges  $Q_1$  and  $Q_2$  are of the same sign, the force is repulsive, if they are of the opposite sign it is attractive. The strength of the electric force decreases as  $1/r^2$  just like the gravitational force between two masses.



Figure 1 Two particles of charge  $Q_1$  and  $Q_2$ , separated by a distance r.

## **CGS Units**

In the CGS system of units, the constant K in Equation (1) is taken to have the numerical value 1, so that Coulomb's law becomes

$$|\vec{F}_{e}| = \frac{Q_{1}Q_{2}}{r^{2}}$$
 (2)

Equation (2) can be used as an experimental definition of charge. Let  $Q_1$  be some accepted standard charge. Then any other charge  $Q_2$  can be determined in terms of the standard  $Q_1$  by measuring the force on  $Q_2$  when the separation is r. In our discussion in Chapter 18, we took the standard  $Q_1$  as the charge on an electron.

This process of defining a standard charge  $Q_1$  and using Equation (1) to determine other charges, is easy in principle but almost impossible in practice. These socalled *electrostatic* measurements are subject to all sorts of experimental problems such as charge leaking away due to a humid atmosphere, redistribution of charge, static charge on the experimenter, etc. Charles Coulomb worked hard just to show that the electric force between two charges did indeed drop as  $1/r^2$ . As a practical matter, Equation (2) is not used to define electric charge. As we will see, the more easily controlled magnetic forces are used instead.

## **MKS Units**

When you buy a 100 *watt* bulb at the store for use in your home, you will see that it is rated for use at 110– 120 *volts* if you live in the United States or Canada, or 220–240 volts most elsewhere. The circuit breakers in your house may allow each circuit to carry up to 15 or 20 *amperes* of current in each circuit. The familiar quantities *volts*, *amperes*, *watts* are all MKS units. The corresponding quantities in CGS units are the totally unfamiliar *statvolts*, *statamps*, and *ergs per second*. Some scientific disciplines, particularly plasma and solid state physics, are conventionally done in CGS units but the rest of the world uses MKS units for describing electrical phenomena. Although we work with the familiar quantities volts, amps and watts using MKS units, there is a price we have to pay for this convenience. In MKS units the constant K in Coulomb's law is written in a rather peculiar way, namely

$$K = \frac{1}{4\pi\epsilon_0}$$
(3)  
$$\epsilon_0 = 8.85 \times 10^{-12} \text{ farads/meter}$$

and Coulomb's law is written as

$$|\vec{F}_{e}| = \frac{Q_1 Q_2}{4\pi\epsilon_0 r^2}$$
  $\overset{Q_1}{\leftarrow}$   $\overset{Q_2}{\leftarrow}$  (4)

where  $Q_1$  and  $Q_2$  are the charges measured in coulombs, and r is the separation measured in meters, and the force  $|\vec{F}_e|$  is in newtons.

Before we can use Equation (4), we have to know how big a unit of charge a coulomb is, and we would probably like to know why there is a  $4\pi$  in the formula, and why the proportionality constant  $\varepsilon_0$  ("*epsilon naught*") is in the denominator.

As we saw in Chapter 18, nature has a basic unit of charge (e) which we call the charge on the electron. A coulomb of charge is  $6.25 \times 10^{18}$  times larger. Just as a liter of water is a large convenient collection of water molecules,  $3.34 \times 10^{25}$  of them, the coulomb can be thought of as a large collection of electron charges,  $6.25 \times 10^{18}$  of them.

In practice, the coulomb is defined experimentally, not by counting electrons, and not by the use of Coulomb's law, but, as we said, by a magnetic force measurement to be described later. For now, just think of the coulomb as a convenient unit made up of  $6.25 \times 10^{18}$  electron charges.

Once the size of the unit charge is chosen, the proportionality constant in Equation (4) can be determined by experiment. If we insist on putting the proportionality constant in the denominator and including a  $4\pi$ , then  $\varepsilon_0$  has the value of  $8.85 \times 10^{-12}$ , which we will often approximate as  $9 \times 10^{-12}$ .

Why is the proportionality constant  $\varepsilon_0$  placed in the denominator? The kindest answer is to say that this is a historical choice that we still have with us. And why include the  $4\pi$ ? There is a better answer to this question. By putting the  $4\pi$  now, we get rid of it in another law that we will discuss shortly, called *Gauss' law*. If you work with Gauss' law, it is convenient to have the  $4\pi$  buried in Coulomb's law. But if you work with Coulomb's law, you will find the  $4\pi$  to be a nuisance.

## **Checking Units in MKS Calculations**

If we write out the units in Equation (4), we get

$$\vec{F}(\text{newtons}) = \frac{Q_1(\text{coul})Q_2(\text{coul})}{4\pi\epsilon_0 r^2(\text{meter})^2}$$
(4a)

In order for the units in Equation (3a) to balance, the proportionality constant  $\varepsilon_0$  must have the dimensions

$$\varepsilon_0 \frac{\text{coulombs}^2}{\text{meter}^2 \text{newton}}$$
 (5)

In earlier work with projectiles, etc., it was often useful to keep track of your units during a calculation as a check for errors. In MKS electrical calculations, it is almost impossible to do so. Units like  $coul^{2/}(meter^{2}newton)$  are bad enough as they are. But if you look up  $\varepsilon_{0}$  in a textbook, you will find its units are listed as *farads/meter*. In other words the combination  $coul^{2/}(newton meter)$  was given the name *farad*. With naming like this, you do not stand a chance of keeping units straight during a calculation. You have to do the best you can to avoid mistakes without having the reassurance that your units check.

#### Summary

The situation with Coulomb's law is not really that bad. We have a  $1/r^2$  force law like gravity, charge is measured in coulombs, which is no worse than measuring mass in kilograms, and the proportionality constant just happens, for historical reasons, to be written as  $1/4\pi\epsilon_0$ .

The units are incomprehensible, so do not worry too much about keeping track of units. After a bit of practice, Coulomb's law will become quite natural.

#### **Example 1** Two Charges

Two positive charges, each 1 coulomb in size, are placed 1 meter apart. What is the electric force between them?



**Solution:** The force will be repulsive, and have a magnitude

$$|\vec{F}_{e}| = \frac{Q_1 Q_2}{4\pi\varepsilon_0 r^2} = \frac{1}{4\pi\varepsilon_0}$$
$$= \frac{1}{4\pi\times9\times10^{-12}} = 10^{10} \text{ newtons}$$

From the answer,  $10^{10}$  newtons, we see that a coulomb is a huge amount of charge. We would not be able to assemble two 1 coulomb charges and put them in the same room. They would tear the room apart.

## **Example 2** Hydrogen Atom

In a classical model of a hydrogen atom, we have a proton at the center of the atom and an electron traveling in a circular orbit around the proton. If the radius of the electron's orbit is  $r = .5 \times 10^{-10}$  meters, how long does it take the electron to go around the proton once?



**Solution:** This problem is more conveniently handled in CGS units, but there is nothing wrong with using MKS units. The charge on the proton is (+e), on the electron (–e), thus the electrical force is attractive and has a magnitude

$$\left|\vec{F}_{e}\right| = \frac{(e)(e)}{4\pi\varepsilon_{0}r^{2}} = \frac{e^{2}}{4\pi\varepsilon_{0}r^{2}}$$

With  $e = 1.6 \times 10^{-19}$  coulombs and  $r = .5 \times 10^{-10}$ m we get

$$\left| \vec{\mathbf{F}_{e}} \right| = \frac{\left( 1.6 \times 10^{-19} \right)^{2}}{4\pi \times \left( 9 \times 10^{-12} \right) \times \left( .5 \times 10^{-10} \right)^{2}}$$
$$= 9 \times 10^{-8} \text{ newtons}$$

Since the electron is in a circular orbit, its acceleration is  $v^2/r$  pointing toward the center of the circle, and we get

$$\left|\vec{a}\right| = \frac{v^2}{r} = \frac{\left|\vec{F}\right|}{m} = \frac{F_e}{m}$$

With the electron mass m equal to  $9.11 \times 10^{-31} \, \text{kg}$  , we have

$$v^{2} = \frac{rF_{e}}{m} = \frac{(.5 \times 10^{-10}) \times 9 \times 10^{-8}}{9.11 \times 10^{-31} \text{ kg}}$$
$$= 4.9 \times 10^{12} \frac{\text{m}^{2}}{\text{s}^{2}}$$
$$v = 2.2 \times 10^{6} \frac{\text{m}}{\text{s}}$$

To go around a circle of radius r at a speed v takes a time

$$T = \frac{2\pi r}{v} = \frac{2\pi \times .5 \times 10^{-10}}{2.2 \times 10^{6}}$$
  
= 1.4×10<sup>-16</sup> seconds

In this calculation, we had to deal with a lot of very small or large numbers, and there was not much of an extra burden putting the  $1/4\pi\epsilon_0$  in Coulomb's law.

#### Exercise 1



(a) Equal numbers of electrons are added to both the earth and the moon until the repulsive electric force exactly balances the attractive gravitational force. How many electrons are added to the earth and what is their total charge in coulombs?

(b) What is the mass, in kilograms of the electrons added to the earth in part (a)?

#### Exercise 2

Calculate the ratio of the electric to the gravitational force between two electrons. Why does your answer not depend upon how far apart the electrons are?

## Exercise 3



Imagine that we could strip all the electrons out of two garden peas, and then placed the peas one meter apart. What would be the repulsive force between them? Express your answer in newtons, and metric tons. (One metric ton is the weight of 1000 kilograms.) (Assume the peas each have about one Avogadro's number, or gram, of protons.) **Exercise 4** 



Two styrofoam balls covered by aluminum foil are suspended by equal length threads from a common point as shown. They are both charged negatively by touching them with a rubber rod that has been rubbed by cat fur. They spread apart by an angle  $2\theta$  as shown. Assuming that an equal amount of charge Q has been placed on each ball, calculate Q if the thread length is  $\ell = 40$  cm, the mass m of the balls is m=10 gm, and the angle is  $\theta = 5^{\circ}$ . Use Coulomb's law in the form  $|\vec{F}_e| = Q_1 Q_2 / 4\pi \epsilon_0 r^2$ , and remember that you must use MKS units for this form of the force law.

## FORCE PRODUCED BY A LINE CHARGE

In our discussion of gravitational forces, we dealt only with point masses because most practical problems deal with spherical objects like moons, planets and stars which can be treated as point masses, or spacecraft which are essentially points.

The kind of problem we did not consider is the following. Suppose an advanced civilization constructed a rod shaped planet shown in Figure (2) that was 200,000 kilometers long and had a radius of 10,000 km. A satellite is launched in a circular orbit of radius 20,000 km, what is the period of the satellite's orbit?

We did not have problems like this because no one has thought of a good reason for constructing a rod shaped planet. Our spherical planets, which can be treated as a point mass, serve well enough.

In studying electrical phenomena, we are not restricted to spherical or point charges. It is easy to spread an electric charge along a rod, and one might want to know what force this charged rod exerted on a nearby point charge. In electricity theory we have to deal with various distributions of electric charge, not just the simple point concentrations we saw in gravitational calculations.



Figure 2

Imagine that an advanced civilization creates a rod-shaped planet. The problem, which we have not encountered earlier, is to calculate the period of a satellite orbiting the planet.

We will see that there is a powerful theorem, discovered by Frederick Gauss, that considerably simplifies the calculation of electric forces produced by extended distribution of charges. But Gauss's law involves several new concepts that we will have to develop. To appreciate this effort, to see why we want to use Gauss's law, we will now do a brute force calculation, using standard calculus steps to calculate the force between a point charge and a line of charge. It will be hard work. Later we will use Gauss's law to do the same calculation and you will see how much easier it is.

The setup for our calculation is shown in Figure (3). We have a negative charge  $Q_T$  located a distance r from a long charged rod as shown. The rod has a positive charge density  $\lambda$  coulombs per meter spread along it. We wish to calculate the total force  $\vec{F}$  exerted by the rod upon our negative test particle  $Q_T$ . For simplicity we may assume that the ends of the rod are infinitely far away (at least several feet away on the scale of the drawing).

To calculate the electric force exerted by the rod, we will conceptually break the rod into many short segments of length dx, each containing an amount of charge dq =  $\lambda dx$ . To preserve the left/right symmetry, we will calculate the force between Q<sub>T</sub> and pairs of dq, one on the left and one on the right as shown. The left directional force  $d\vec{F}_1$  and the right directed force  $d\vec{F}_2$  add up to produce an upward directional force  $d\vec{F}$ . Thus, when we add up the forces from pairs of dq, the  $d\vec{F}$ 's are all upward directed and add numerically.

The force  $d\vec{F}_1$  between  $Q_T$  and the piece of charge  $dq_1$  is given by Coulomb's law as

$$|d\vec{F}_{1}| = \frac{KQ_{T}dq_{1}}{R^{2}} = \frac{KQ_{T}dq_{1}}{x^{2} + r^{2}}$$
$$= \frac{K\lambda Q_{T}dx}{x^{2} + r^{2}}$$
(6)

where, for now, we will use K for  $1/4\pi\epsilon_0$  to simplify the formula. The separation R between  $Q_T$  and  $dq_1$  is given by the Pythagorean theorem as  $R^2 = x^2 + r^2$ .

The component of  $\vec{dF_1}$  in the upward direction is  $dF_1 \cos(\theta)$ , so that  $\vec{dF}$  has a magnitude

$$\left| d\vec{F} \right| = 2 dF_1 \cos(\theta)$$
$$= 2 \frac{K\lambda Q_T dx}{x^2 + r^2} \frac{r}{\sqrt{x^2 + r^2}}$$
(7)

The factor of 2 comes from the fact that we get equal components from both  $d\vec{F}_1$  and  $d\vec{F}_2$ .

To get the total force on  $Q_T$ , we add up the forces produced by all pairs of dq starting from x = 0 and going out to  $x = \infty$ . The result is the definite integral

$$\left|\vec{F}_{Q_{T}}\right| = \int_{0}^{\infty} \frac{2K\lambda Q_{T}r}{(x^{2}+r^{2})^{3/2}} dx$$
$$= 2K\lambda Q_{T}r \int_{0}^{\infty} \frac{dx}{(x^{2}+r^{2})^{3/2}}$$
(8)

where r, the distance from the charge to the rod, is a constant that can be taken outside the integral.

The remaining integral  $\int dx/(x^2 + r^2)^{3/2}$  is not a common integral whose result you are likely to have memorized, nor is it particularly easy to work out.

Instead we look it up in a table of integrals with the result

$$\int_{0}^{A} \frac{dx}{(x^{2} + r^{2})^{\frac{3}{2}}} = \frac{2(2x)}{4r^{2}\sqrt{r^{2} + x^{2}}} \Big|_{0}^{A}$$
$$= \frac{A}{r^{2}\sqrt{r^{2} + A^{2}}} - 0$$
(9)

For A>>r (very long rod), we can set  $\sqrt{r^2 + A^2} \approx A$ and we get the results

$$\int_{0}^{A >>r} \frac{dx}{(x^{2} + r^{2})^{\frac{3}{2}}} = \frac{A}{r^{2}A} = \frac{1}{r^{2}}$$
(10)

Using Equation (10) in Equation (8) we get

$$\left| \vec{F}_{Q_{T}} \right| = 2K\lambda Q_{T} \frac{r}{r^{2}}$$

$$F_{Q_{T}} = \frac{2K\lambda Q_{T}}{r}$$
(11)

The important point of the calculation is that *the force between a point charge and a line charge drops off as* 1/r *rather than*  $1/r^2$ , as long as  $Q_T$  stays close enough to the rod that the ends appear to be very far away.



#### Figure 3

Geometry for calculating the electric force between a point charge  $Q_T$  a distance r from a line of charge with  $\lambda$  coulombs per meter.

One of the rules of thumb in doing physics is that if you have a simple result, there is probably an intuitive derivation or explanation. In deriving the answer in Equation (11), we did too much busy work to see anything intuitive. We had to deal with an integral of  $(x^2 + r^2)^{-3/2}$ , yet we got the simple answer that the force dropped off as 1/r rather than the  $1/r^2$ . We will see, when we repeat this derivation using Gauss's law, that the change from  $1/r^2$  to a 1/r force results from the change from a three to a two dimensional problem. This basic connection with geometry is not obvious in our brute force derivation.

#### Exercise 5

Back to science fiction. A rod shaped planet has a mass density  $\lambda$  kilograms per unit length as shown in Figure (4). A satellite of mass m is located a distance r from the rod as shown. Find the magnitude of the gravitational force  $\vec{F}_g$  exerted on the satellite by the rod shaped planet. Then find a formula for the period of the satellite in a circular orbit.



**Figure 4** *Rod shaped planet and satellite.* 

## Short Rod

Our brute force calculation does have one advantage, however. If we change the problem and say that our charge is located a distance r from the center of a finite rod of length 2L as shown in Figure (5), then Equation (8) of our earlier derivation becomes

$$F_{Q_{T}} = 2K\lambda Q_{T}r \int_{0}^{L} \frac{dx}{(x^{2} + r^{2})^{\frac{3}{2}}}$$
(12)

The only difference is that the integral stops at L rather than going out to infinity.

From Equation (9) we have

$$\int_{0}^{L} \frac{dx}{(x^{2} + r^{2})^{\frac{3}{2}}} = \frac{L}{r^{2}\sqrt{r^{2} + L^{2}}}$$
(13)

And the formula for the force on Q<sub>T</sub> becomes



A harder problem is to calculate the force Fexerted on  $Q_T$  by a rod of finite length 2 L.

Equation (14) has the advantage that it can handle both limiting cases of a long rod (L>>r), a short rod or point charge (L << r), or anything in between.

For example, if we are far away from a short rod, so that  $L \ll r$ , then  $\sqrt{r^2 + L^2} \approx r$ . Using the fact that  $2L\lambda = Q_R$  is the total charge on the rod, Equation (14) becomes

$$F_{Q_T} = \frac{KQ_RQ_T}{r^2} \quad (L << r) \tag{15b}$$

which is just Coulomb's law for point charges. The more general result, Equation (14), which we obtained by the brute force calculation, cannot be obtained with simple arguments using Gauss's law. This formula was worth the effort.

#### **Exercise 6**

Show that if we are very close to the rod, i.e.  $r \ll L$ , then Equation (14) becomes the formula for the force exerted by a line charge.

## THE ELECTRIC FIELD

Our example of the force between a point charge and a line of charge demonstrates that even for simple distributions of charge, the calculation of electric forces can become complex. We now begin the introduction of several new concepts that will allow us to simplify many of these calculations. The first of these is the *electric field*, a concept which allows us to rely more on maps, pictures and intuition, than upon formal calculations.

To introduce the idea of an electric field, let us start with the simple distribution of charge shown in Figure (6). A positive charge of magnitude  $Q_A$  is located at Point A, and a negative charge of magnitude  $-Q_B$  is located at Point B. We will assume that these charges  $Q_A$  and  $Q_B$  are fixed, nailed down. They are our *fixed charge distribution*.

We also have a positive *test charge* of magnitude  $Q_T$  that we can move around in the space surrounding the fixed charges.  $Q_T$  will be used to test the strength of the electric force at various points, thus the name "test charge".

In Figure (6), we see that the test charge is subject to the repulsive force  $\vec{F}_A$  and attractive force  $\vec{F}_B$  to give a net force  $\vec{F} = \vec{F}_A + \vec{F}_B$ . The individual forces  $\vec{F}_A$  and  $\vec{F}_B$  are given by Coulomb's law as



where  $r_A$  is the distance from  $Q_T$  to  $Q_A$  and  $r_B$  is the distance from  $Q_T$  to  $Q_B$ . For now we are writing  $1/4\pi\epsilon_0 = K$  to keep the formulas from looking too messy.

In Figure (7), we have the same distribution of fixed charge as in Figure (6), namely  $Q_A$  and  $Q_B$ , but we are using a smaller test charge  $Q'_T$ . For this sketch,  $Q'_T$  is about half as big as  $Q_T$  of Figure (6), and the resulting force vectors point in the same directions but are about half as long;

$$\left|\vec{F'}_{A}\right| = \frac{KQ'_{T}Q_{A}}{r_{A}^{2}}$$

$$\left|\vec{F'}_{B}\right| = \frac{KQ'_{T}Q_{B}}{r_{B}^{2}}$$
(17)

The only difference between Equations (16) and (17) is that  $Q_T$  has been replaced by  $Q'_T$  in the formulas. You can see that Equations (17) can be obtained from Equations (16) by multiplying the forces by  $Q'_T/Q_T$ . Thus if we used a standard size test charge  $Q_T$  to calculate the forces, i.e. do all the vector additions, etc., then we can find the force on a different sized test charge  $Q'_T$  by multiplying the net force by the ratio  $Q'_T/Q_T$ .



**Figure 6** Forces exerted by two fixed charges  $Q_A$  and  $Q_B$  on the test particle  $Q_T$ .







#### **Figure 7** If we replace th

If we replace the test charge  $Q_T$  by a smaller test charge  $Q_T$ , everything is the same except that the force vectors become shorter.

#### Unit Test Charge

The next step is to decide what size our standard test charge  $Q_T$  should be. Physically  $Q_T$  should be small so that it does not disturb the fixed distribution of charge. After all, Newton's third law requires that  $Q_T$  pull on the fixed charges with forces equal and opposite to the forces shown acting on  $Q_T$ .

On the other hand a simple mathematical choice is  $Q_T = 1$  coulomb, what we will call a "*unittest charge*". If  $Q_T = 1$ , then the force on another charge  $Q'_T$  is just  $Q'_T$  times larger  $(\vec{F} = \vec{F}Q'_T/Q_T = \vec{F}Q'_T/1 = \vec{F}Q'_T)$ . The problem is that in practice, a coulomb of charge is enormous. Two point charges, each of strength +1 coulomb, located one meter apart, repel each other with a force of magnitude

$$\left| \vec{F} \right| = \frac{K \times 1 \text{ coulomb} \times 1 \text{ coulomb}}{1^2 \text{m}^2}$$
  
=  $K = \frac{1}{4\pi\epsilon_0} = 9 \times 10^9 \text{ newtons}$  (17)

a result we saw in Example 1. A force of nearly 10 billion newtons is strong enough to destroy any experimental structure you are ever likely to see. In practice the coulomb is much too big a charge to serve as a realistic test particle.

The mathematical simplicity of using a unit test charge is too great to ignore. Our compromise is the following. We use a unit test charge, but think of it as a "small unit test charge". Conceptually think of using a charge about the size of the charge of an electron, so that the



#### Figure 8

When we use a unit test charge  $Q_T = 1$  coulomb, then the forces on it are called "electric field" vectors,  $\vec{E}_A$ ,  $\vec{E}_B$  and  $\vec{E}$  as shown.

forces it exerts do not disturb anything, but mathematically treat the test charge as having a magnitude of 1 coulomb. In addition, we will always use a *positive unit* test charge. If we want to know the force on a negative test charge, simply reverse the direction of the force vectors.

Figure (8) is the same as our Figures (6) and (7), except that we are now using a unit test particle equal to 1 coulomb, to observe the electric forces surrounding our fixed charge distribution of  $Q_A$  and  $Q_B$ . The forces acting on  $Q_T = 1$  coulomb are

$$\left|\vec{F}_{A}\right| = \frac{KQ_{A} \times 1 \text{ coulomb}}{r_{A}^{2}}$$

$$\left|\vec{F}_{B}\right| = \frac{KQ_{B} \times 1 \text{ coulomb}}{r_{B}^{2}}$$
(18)

To emphasize that these forces are acting on a unit test charge, we will use the letter E rather than F, and write

$$\left|\vec{E}_{A}\right| = \frac{KQ_{A}}{r_{A}^{2}}$$

$$\left|\vec{E}_{B}\right| = \frac{KQ_{B}}{r_{B}^{2}}$$
(19)

If you wished to know the force  $\vec{F}$  on some charge Q located where our test particle is, you would write

$$\left|\vec{F}_{A}\right| = \frac{KQ_{A}Q}{r_{A}^{2}} = \left|\vec{E}_{A}\right|Q$$
 (21a)

$$\left|\vec{F}_{B}\right| = \frac{KQ_{B}Q}{r_{B}^{2}} = \left|\vec{E}_{B}\right|Q$$
(21b)

$$\vec{F} = Q\vec{E}_{A} + Q\vec{E}_{B} = Q(\vec{E}_{A} + \vec{E}_{B})$$
$$\vec{F} = Q\vec{E}$$
(22)

where  $\vec{E} = \vec{E}_A + \vec{E}_B$ . Equation (22) is an important result. It says that *the force on any charge Q is Q times the force*  $\vec{E}$  *on a unit test particle.* 

## **ELECTRIC FIELD LINES**

The force  $\vec{E}$  on a unit test particle plays such a central role in the theory of electricity that we give it a special name – the *electric field*  $\vec{E}$ .

$$\begin{array}{c} \text{electric} \\ \text{field} \vec{E} \end{array} \right\} \equiv \begin{cases} \text{force on a} \\ \text{unit test} \\ \text{particle} \end{cases}$$

(23)

Once we know the electric field  $\vec{E}$  at some point, then the force  $\vec{F}_Q$  on a charge Q located at that point is

$$\vec{F} = Q\vec{E}$$
(24)

If Q is negative, then  $\vec{F}$  points in the direction opposite to  $\vec{E}$ .

From Equation (24), we see that the electric field  $\vec{E}$  has the dimensions of newtons/coulomb, so that  $Q\vec{E}$  comes out in newtons.



#### Figure 9

Once we know the electric field  $\vec{E}$  at some point, we find the force  $\vec{F}$  acting on a charge Q at that point by the simple formula  $\vec{F} = Q\vec{E}$ .

## **Mapping the Electric Field**

In Figure (10), we started with a simple charge distribution +Q and -Q as shown, placed our unit test particle at various points in the region surrounding the fixed charges, and drew the resulting force vectors  $\vec{E}$  at each point. If we do the diagram carefully, as in Figure (10), a picture of the electric field begins to emerge. Once we have a complete picture of the electric field  $\vec{E}$ , once we know  $\vec{E}$  at every point in space, then we can find the force on any charge q by using  $\vec{F}_Q = Q\vec{E}$ . The problem we wish to solve, therefore, is how to construct a complete map or picture of the electric field  $\vec{E}$ .

#### **Exercise 7**

In Figure (10), we have labeled 3 points (1), (2), and (3). Sketch the force vectors  $\vec{F}_Q$  on :

- (a) a charge Q = 1 coulomb at Point (1)
- (b) a charge Q = -1 coulomb at Point (2)
- (c) a charge Q = 2 coulombs at Point (3)



#### Figure 10

If we draw the electric field vectors  $\overline{E}$  at various points around our charge distribution, a picture or map of the electric field begins to emerge.

In part (c) of Exercise (7), we asked you to sketch the force on a charge Q = 2 coulombs located at Point (3). The answer is  $\vec{F}_3 = Q\vec{E}_3 = 2\vec{E}_3$ , but the problem is that we have not yet calculated the electric field  $\vec{E}_3$  at Point (3). On the other hand we have calculated  $\vec{E}$  at some nearby locations. From the shape of the map that is emerging from the  $\vec{E}$  vectors we have drawn, we can make a fairly accurate guess as to the magnitude and direction of  $\vec{E}$  at Point (3) without doing the calculation. With a map we can build intuition and make reasonably accurate estimates without calculating  $\vec{E}$  at every point.

In Figure (10) we were quite careful about choosing where to draw the vectors in order to construct the picture. We placed the points one after another to see the flow of the field from the positive to the negative charge. In Figure (11), we have constructed a similar picture for the electric field surrounding a single positive charge.

The difficulty in drawing maps or pictures of the electric field is that we have to show both the magnitude and direction at every point. To do this by drawing a large number of separate vectors quickly becomes cumbersome and time consuming. We need a better way to draw these maps, and in so doing will adopt many of the conventions developed by map makers.



Electric field of a point charge.

## **Field Lines**

As a first step in simplifying the mapping process, let us concentrate on showing the direction of the electric force in the space surrounding our charge distribution. This can be done by connecting the arrows in Figures (10) and (11) to produce the line drawings of Figures (12a) and (12b) respectively. The lines in these drawings are called *field lines*.

In our earlier discussion of fluid flow we saw diagrams that looked very much like the Figures (12a and 12b). There we were drawing stream lines for various flow patterns. Now we are drawing electric field lines. As illustrated in Figure (13), a streamline and an electric field line are similar concepts. At every point on a streamline, the velocity field  $\vec{v}$  is parallel to the streamline, while at every point on an electric field line, the electric field  $\vec{E}$  is parallel to the electric field line.



**Figure 12a** We connected the arrows of Figure 10 to create a set of field lines for 2 point charges.



**Figure 12b** Here we connected the arrows in Figure 11 to draw the field lines for a point charge.

## **Continuity Equation for Electric Fields**

Figure (14) is our old diagram (23-6) for the velocity field of a point source of fluid (a small sphere that created water molecules). We applied the continuity equation to the flow outside the source and saw that the velocity field of a point source of fluid drops off as  $1/r^2$ .

Figure (15) is more or less a repeat of Figure (12b) for the electric field of a point charge. By Coulomb's law, the strength of the electric field drops off as  $1/r^2$  as we go out from the point charge. We have the same field structure for a point source of an incompressible fluid and the electric field of a point charge. Is this pure coincidence, or is there something we can learn from the similarity of these two fields?

The crucial feature of the velocity field that gave us a  $1/r^2$  flow was the continuity equation. Basically the idea is that all of the water that is created in the small sphere must eventually flow out through any larger sphere surrounding the source. Since the area of a sphere,  $4\pi r^2$ , increases as  $r^2$ , the speed of the water has to decrease as  $1/r^2$  so that the same volume of water per second flows through a big sphere as through a small



Figure 13

Comparison of the streamline for the velocity field

and the field line for an electric field. Both are constructed in the same way by connecting successive vectors. The streamline is easier to visualize because it is the actual path followed by particles in the fluid.

If we think of the electric field as some kind of an incompressible fluid, and think of a point charge as a source of this fluid, then the continuity equation applied to this electric field gives us the correct  $1/r^2$ dependence of the field. In a sense we can replace Coulomb's law by a continuity equation. Explicitly, we will use streamlines or field lines to map the direction of the field, and use the continuity equation to calculate the magnitude of the field. This is our general plan for constructing electric field maps; we now have to fill in the details.



#### Figure 14

This is our old Figure 23-6 for the velocity field of a point source of water. The continuity equation  $v_1A_1 = v_2A_2$ , requires that the velocity field  $\vec{v}$  drops off as  $1/r^2$  because the area through which the water flows increases as  $\pi r^2$ .



#### Figure 15

From Coulomb's law,  $|\vec{E}| = KQ/r^2$ , we see that the electric field of a point source drops off in exactly the same way as the velocity field of a point source. Thus the electric field must obey the same continuity equation  $E_1A_1 = E_2A_2$  as does the velocity field.

## Flux

To see how the continuity equation can be applied to electric fields, let us review the calculation of the  $1/r^2$  velocity field of a point source of fluid, and follow the same steps to calculate the electric field of a point charge.

In Figure (16) we have a small sphere of area  $A_1$  in which the water is created. The volume of water created each second, which we called the *flux* of the water and will now designate by the greek letter  $\Phi$ , is given by

volume of water  
created per second  
in the small sphere 
$$= \Phi_1 = v_1 A_1$$
 (25)

The flux of water out through a larger sphere of area  $A_2$  is



#### Figure 16

The total flux  $\Phi_1$  of water out of the small sphere is  $\Phi_1 = v_1 A_{1\perp}$ , where  $A_{1\perp}$  is the perpendicular area through which the water flows. The flux through the larger sphere is  $\Phi_2 = v_2 A_{2\perp}$ . Noting that no water is lost as it flows from the inner to outer sphere, i.e., equating  $\Phi_1$  and  $\Phi_2$ , gives us the result that the velocity field drops off as  $1/r^2$  because the perpendicular area increases as  $r^2$ .

The continuity equation  $v_1A_1 = v_2A_2$  requires these fluxes be equal

$$\Phi_1 = \Phi_2 \equiv \Phi \qquad \begin{array}{c} continuity \\ equation \end{array} \tag{27}$$

Using Equations (26) and (27), we can express the velocity field  $v_2$  out at the larger sphere in terms of the flux of water  $\Phi$  created inside the small sphere

$$v_2 = \frac{\Phi}{A_2} = \frac{\Phi}{4\pi r_2^2}$$
(28)

Let us now follow precisely the same steps for the electric field of a point charge. Construct a small sphere of area  $A_1$  and a large sphere  $A_2$  concentrically surrounding the point charge as shown in Figure (17). At the small sphere the electric field has a strength  $E_1$ , which has dropped to a strength  $E_2$  out at  $A_2$ .

Let us define  $E_1A_1$  as the flux of our electric fluid flowing out of the smaller sphere, and  $E_2A_2$  as the flux flowing out through the larger sphere

$$\Phi_1 = \mathbf{E}_1 \mathbf{A}_1 \tag{29}$$



#### Figure 17

The total flux  $\Phi_1$  of the electric field out of the small sphere is  $\Phi_1 = E_1 A_{1\perp}$ , where  $A_{1\perp}$  is the perpendicular area through which the electric field flows. The flux through the larger sphere is  $\Phi_2 = E_2 A_{2\perp}$ . Noting that no flux is lost as it flows from the inner to outer sphere, i.e., equating  $\Phi_1$ and  $\Phi_2$ , gives us the result that the electric field drops off as  $1/r^2$  because the perpendicular area increases as  $r^2$ .

Applying the continuity equation  $E_1A_1 = E_2A_2$  to this electric fluid, we get

$$\Phi_1 = \Phi_2 = \Phi \tag{31}$$

Again we can express the field at A<sub>2</sub> in terms of the flux

$$E_2 = \frac{\Phi}{A_2} = \frac{\Phi}{4\pi r_2^2} \tag{32}$$

Since  $A_2$  can be any sphere outside, but centered on the point charge, we can drop the subscript 2 and write

$$E(r) = \frac{\Phi}{4\pi r^2}$$
(33)

In Equation (33), we got the correct  $1/r^2$  dependence for the electric field, but what is the appropriate value for  $\Phi$ ? How much electric flux  $\Phi$  flows out of a point charge?

To find out, start with a fixed charge Q as shown in Figure (18), place our unit test charge a distance r away, and use Coulomb's law to calculate the electric force E on our unit test charge. The result is



#### Figure 18

Using Coulomb's law for the electric field of a point charge Q, we calculate that the total flux out through any centered sphere surrounding the point charge is  $\Phi = Q/\varepsilon_0$ .

$$\left| \vec{E} \right| = \frac{Q}{4\pi\epsilon_0 r^2}$$
 Coulombs (34)

where now we are explicitly putting in  $1/4\pi\varepsilon_0$  for the proportionality constant K.

Comparing Equations (33) and (34), we see that if we choose

$$\Phi = \frac{Q}{\varepsilon_{o}}$$
Flux emerging  
from a charge Q (35)

then the continuity equation (33) and Coulomb's law (34) give the same answer. Equation (35) is the key that allows us to apply the continuity equation to the electric field. If we say that a point charge Q creates an electric flux  $\Phi = Q/\epsilon_o$ , then applying the continuity equation gives the same results as Coulomb's law. (You can now see that by putting the  $4\pi$  into Coulomb's law, there is no  $4\pi$  in our formula (35) for flux.)

#### **Negative Charge**

If we have a negative charge -Q, then our unit test particle  $Q_T$  will be attracted to it as shown in Figure (19). From a hydrodynamic point of view, the electric fluid is flowing **into** the charge -Q and being destroyed there. Therefore a generalization of our rule about electric flux is that a positive charge creates a positive, outward flux of magnitude  $Q/\varepsilon_o$ , while a negative charge destroys the electric flux, it has a negative flux  $- Q/\varepsilon_o$  that flows into the charge and disappears.




## **Flux Tubes**

In our first pictures of fluid flows like Figure (23-5) reproduced here, we saw that the streamlines were little tubes of flow. The continuity equation, applied to a streamline was  $\Phi = v_1A_1 = v_2A_2$ . This is simply the statement that the flux of a fluid along a streamline is constant. We can think of the streamlines as small tubes of flux. By analogy *we will think of our electric field lines as small tubes of electric flux*.

## **Conserved Field Lines**

When we think of the field line as a small flux tube, the continuity equation gives us a very powerful result, namely the flux tubes must be continuous, must maintain their strength in any region where the fluid is neither being created nor destroyed. For the electric fluid, the flux tubes or field lines are created by, start at, positive charge. And they are destroyed by, or stop at, negative charge. But in between the electric fluid is conserved and the field lines are continuous. We will see that this continuity of the electric field lines is a very powerful tool for mapping electric fields.



**Figure 23-5** *Flux tube in the flow of water.* 

#### A Mapping Convention

If an electric field line represents a small flux tube, the question remains as to how much flux is in the tube? Just as we standardized on a unit test charge  $Q_T = 1$  coulomb for the definition of the electric field  $\vec{E}$ , we will standardize on a unit flux tube as the amount of flux represented by one electric field line. With this convention, we should therefore draw  $\Phi = Q/\epsilon_0$  field lines or unit flux tubes coming out of a positive charge +Q, or stopping on a negative charge –Q. Let us try a few examples to see what a powerful mapping convention this is.

In Figure (20) we have a positive charge  $Q/\varepsilon_0 = +5$ , and a negative charge  $Q/\varepsilon_0 = -3$ , located as shown. By our new mapping convention we should draw 5 unit flux tubes or field lines out of the positive charge, and we should show 3 of them stopping on the negative charge. Close to the positive charge, the negative charge is too far away to have any effect and the field lines must go radially out as shown. Close to the negative charge, the lines must go radially in because the positive charge is too far away.



#### Figure 20

We begin a sketch of the electric field by drawing the field lines in close to the charges, where the lines go either straight in or straight out. Here we have drawn 3 lines into the charge  $-3\varepsilon_0$  and 5 lines out of the charge  $+5\varepsilon_0$ . To make a symmetric looking picture, we oriented the lines so that one will go straight across from the positive to the negative charge. Now we get to the interesting part; what happens to the field lines out from the charges? The basic rule is that *the lines can start on positive charge, stop on negative charge, but must be continuous in between*. A good guess is that 3 of the lines starting on the positive charge go over to the negative charge in roughly the way we have drawn by dotted lines in Figure (21). There is no more room on the minus charge for the other two lines, so that all these two lines can do is to continue on out to infinity.

Let us take Figure (21), but step far back, so that the + and the – charge look close together as shown in Figure (22). Between the charges we still have the same heartshaped pattern, but we now get a better view of the two lines that had nowhere to go in Figure (21). To get a better understanding for Figure (22) draw a sphere around the charges as shown. The net charge inside this sphere is

$$\left(\frac{Q_{\text{net}}}{\varepsilon_0}\right)_{\text{inside sphere}} = +5-3 = 2$$

Thus by our mapping convention, two field lines should emerge from this sphere, and they do. Far away



Figure 21

Once the in close field lines have been drawn, we can sketch in the connecting part of the lines as shown above. Three of the lines starting from the positive charge must end on the negative one. The other must go out to infinity. Using symmetry and a bit of artistic skill, you will become quite good at drawing these sketches. where we cannot see the space between the point charges, it looks like we have a single positive charge of magnitude  $Q/\varepsilon_0 = 2$ .

## Summary

When we started this chapter with the brute force calculus calculation of the electric field of a line charge, you may have thought that the important point of this chapter was how to do messy calculations. Actually, exactly the opposite is true! We want to learn how to avoid doing messy calculations. The sketches shown in Figures (21) and (22) are an important step in this process. From what we are trying to get out of this chapter, it is far more important that you learn how to do sketches like Figures (21) and (22), than calculations illustrated by Figure (1).

With a little experience, most students get quite good at sketching field patterns. The basic constraints are that  $Q/\epsilon_0$  lines start on positive charges, or stop on negative charges. Between charges the lines are unbroken and should be smooth, and any lines left over must either go to or come from infinity as they did in Figure (22).



#### Figure 22

Distant view of our charge distribution. If we step way back from the charge distribution of Figure (21), we see a small object whose net charge is  $-3\varepsilon_0 + 5\varepsilon_0 = +2\varepsilon_0$ . Thus 2 lines must finally emerge from this distribution as shown.

# A Computer Plot

Figure (23) is a computer plot of the electric field lines for the +5, -3 charge distribution of Figures (21) and (22). The first thing we noted is that the computer drew a lot more lines than we did. Did it violate our mapping convention that the lines represent unit flux tubes, with  $Q/\varepsilon_0$  lines starting or stopping on a charge Q? Yes. The computer drew a whole bunch of lines so that we could get a better feeling for the shape of the electric field. Notice, however, that the ratio of the number of lines starting from the positive charge to the number ending on the negative charge is still 5/3. One of the standard tricks in map making is to change your scale to make the map look as good as possible. Here the computer drew 10 lines per unit flux tube rather than 1. We will see that the only time we really have to be careful with the number of lines we draw is when we are using a count of the number of lines to estimate the strength of the electric field.



## Figure 23

Computer plot of the field lines of a - 3 and + 5 charge distribution. Rather than drawing  $Q/\epsilon_0$  out of a positive charge or into a negative charge, the computer is programmed to draw enough lines to make the shape of the electric field as clear as possible.

## **GAUSS' LAW**

The idea of using the continuity equation to map field lines was invented by Frederick Gauss and is known as *Gauss' law*. A basic statement of the law is as follows. Conceptually construct a closed surface (often called a *Gaussian surface*) around a group of sources or sinks, as shown in Figure (24). In that figure we have drawn the Gaussian surface around three sources and one sink. Then calculate the total flux  $\Phi_{tot}$  coming from these sources. For Figure (24), we have

$$\Phi_{\text{tot}} = \Phi_1 + \Phi_2 + \Phi_3 + \Phi_4 \tag{36}$$

where  $\Phi_2$  happens to be negative. Then if the fluid is incompressible, or we have an electric field, the total flux flowing out through the Gaussian surface must be equal to the amount of flux  $\Phi_{tot}$  being created inside.

Gauss' law applies to any closed surface surrounding our sources and sinks. But the law is useful for calculations when the Gaussian surface is simple enough in shape that we can easily write the formula for the flux flowing through the surface. To illustrate the way we use Gauss' law, let us, one more time, calculate the electric field of a point charge.



#### Figure 24

If these were sources and sinks in a fluid, it would be obvious that the total flux of fluid out through the closed surface is equal to the net amount of fluid created inside. The same concept applies to electric flux. The net flux  $\Phi_{tot}$  out through the Gaussian surface is the sum of the fluxes  $\Phi_1 + \Phi_2 + \Phi_3 + \Phi_4$  created inside. In Figure (25), we have a point charge Q, and have drawn a spherical Gaussian surface around the charge. The flux produced by the point charge is

$$\Phi_{\text{tot}} = \frac{Q}{\varepsilon_0} \qquad \begin{array}{c} flux \ produced \\ by \ the \ point \\ charge \end{array} \tag{37}$$

At the Gaussian surface there is an electric field  $\vec{E}$  (which we wish to calculate), and the surface has an area A =  $4\pi r^2$ . Therefore the electric flux flowing out through the sphere is

$$\Phi_{out} = EA$$

$$= E \times 4\pi r^{2} \qquad \begin{array}{c} flux \ flowing \\ out \ through \\ the \ sphere \end{array}$$
(38)

Equating the flux created inside (Equation 37) to the flux flowing out (Equation 38) gives





Calculating the electric field of a point charge by equating the flux  $Q/\varepsilon_0$  created by the point charge to the flux  $\Phi = EA_{\perp}$  flowing through the Gaussian surface. This gives

$$Q/\varepsilon_0 = E \times 4\pi r^2$$
$$E = Q/4\pi\varepsilon_0 r^2$$
as we expect.

## **Electric Field of a Line Charge**

As a real test of Gauss' law, let us calculate the electric field of a line charge and compare the result with our brute force calculus calculation. In the calculus derivation, we found that the force on a test charge  $Q_T$  a distance r from our line charge was (from Equation (11))

$$\vec{F} = \frac{2K\lambda Q_T}{r} = \frac{2\lambda Q_T}{4\pi\varepsilon_0 r}$$

where  $\lambda$  is the charge density on the rod as shown in Figure (26). Setting  $Q_T = 1$  coulomb,  $\vec{F}$  becomes the force  $\vec{E}$  on a unit test charge:

$$\begin{vmatrix} \vec{E} \end{vmatrix} = \frac{\lambda}{2\pi\epsilon_{o}r} \qquad \begin{array}{c} calculus derived \\ formula for the \\ electric field of \\ a line charge \end{matrix}$$
(40)

To apply Gauss' law we first construct a cylindrical Gaussian surface that surrounds a length L of the charge as shown in Figure (27). Since the charge density is  $\lambda$ , the total charge  $Q_{in}$ , inside our Gaussian surface is

 $Q_{in} = \lambda L$ 

This amount of charge creates an amount of flux

$$\Phi_{\text{inside}} = \frac{Q_{\text{in}}}{\varepsilon_{\text{o}}} = \frac{\lambda L}{\varepsilon_{\text{o}}}$$
(41)

Now we can see from Figure (27) that because the electric field lines go radially outward from the line charge, they only flow out through the curved outer surface of our Gaussian cylinder and not through the flat ends. This cylindrical surface has an area  $A = (2\pi r)L$  (circumference × length), and the electric field out at a distance r is E(r). Thus the flux out through the Gaussian cylinder is

$$\Phi_{\rm out} = E(r) 2\pi r L \tag{42}$$





The force  $\vec{E}$  on a unit test charge near a line charge.

Equating the flux created inside (Equation 41) to the flux flowing out through the cylindrical surface (Equation 42) gives

$$E(r) 2\pi rL = \frac{\lambda L}{\epsilon_0}$$

The L's cancel and we get

$$E(r) = \frac{\lambda}{2\pi\varepsilon_{o}r}$$
(43)

Voilà! We get the same result. Compare the calculus derivation with its integral of  $(x^2 + r^2)^{3/2}$ , to the simple steps of Equations (41) and (42). We noted that in physics, a simple answer, like the 1/r dependence of the electric field of a line charge, should have an easy derivation. The easy derivation is Gauss' law. The simple idea is that for a line charge the flux is flowing out through a cylindrical rather than a spherical surface. The area of a cylindrical surface increases as r, rather than as r<sup>2</sup> for a sphere, therefore the electric field drops off as 1/r rather than as 1/r<sup>2</sup> as it did for a point charge.



#### Figure 27

Using Gauss' law to calculate the electric field of a line charge. Draw the Gaussian surface around a section of the rod. The flux all flows out through the cylindrical surface.

## Flux Calculations

In our calculations of the flux through a Gaussian surface, Equation (38) for a spherical surface and Equation (42) for a cylindrical surface, we multiplied the strength of the electric field times the area A through which the field was flowing. In both cases we were careful to construct the area perpendicular to the field lines, for flux is equal to the strength of the field times cross sectional or perpendicular area through which it is flowing. It would be more accurate to write the formula for flux as

$$\Phi = E A_{\perp} \tag{44}$$

where the  $\perp$  sign reminds us that  $A_{\perp}$  is the perpendicular area.

## Area as a Vector

A more formal way to present the formula for flux is to turn the area A into a vector. To illustrate the procedure, consider the small flow tube shown in Figure (28). We have sliced the tube with a plane, and the intersection of the tube and the plane gives us an area A as shown. We turn A into the vector  $\vec{A}$  by drawing an arrow perpendicular to the plane, and of length A.

To show why we bothered turning A into a vector, in Figure (29) we have constructed a cross-sectional area  $\vec{A}_{\perp}$  as well as the area  $\vec{A}$  of Figure (28). The cross-sectional area is the smallest area we can construct



Figure 28

Definition of the area as a vector. Slice a flow tube by a plane. The area A is the area of the region in the plane bounded by the flow tube. We define the direction of A as pointing perpendicular to the plane as shown. across the tube. The other areas are bigger by a factor  $1/\cos(\theta)$  where  $\theta$  is the angle between  $\vec{A}$  and  $\vec{A}_{\perp}$ . I.e.,

$$A_{\perp} = A\cos(\theta) \tag{45}$$

Now the flux in the flow tube is the fluid speed v times the cross-sectional area  $A_{\perp}$ 

$$\Phi = vA_{\perp} \tag{46}$$

Using Equation (45) for  $A_{\perp}$  in Equation (46) gives

$$\Phi = vA_{\perp} = vA\cos(\theta)$$

But vAcos  $(\theta)$  is just the vector dot product of the velocity vector  $\vec{v}$  (which points in the direction  $\vec{A}_{\perp}$ ) and  $\vec{A}$ , thus we have the more general formula

$$\Phi = vA_{\perp} = \vec{v} \cdot \vec{A} \tag{47a}$$

By analogy, the electric flux through an area  $\vec{A}$  is

$$\Phi = \mathbf{E}\mathbf{A}_{\perp} = \vec{\mathbf{E}} \cdot \vec{\mathbf{A}} \tag{47b}$$

In extreme cases where the Gaussian surface is not smooth, you may have to break up the surface into small pieces, calculate the flux  $d\Phi_i = \vec{E} \cdot d\vec{A}_i$  for each piece  $d\vec{A}_i$  and then add up all the contributions from each piece to get the total flux  $\Phi_{out}$ . The result is called a **surface integral** which we will discuss later. For now we will make sure that our Gaussian surfaces are smooth and perpendicular to the field, so that we can use the simple form of Equation (47).



#### Figure 29

The cross-sectional area, which we have been calling  $\vec{A}_{\perp}$ , is the smallest area that crosses the entire tube. The tilted area is larger than  $\vec{A}_{\perp}$  by a factor of  $1/\cos(\theta)$ , where  $\theta$  is the angle between  $\vec{A}$  and  $\vec{A}_{\perp}$ . As a result we can write

$$A_{\perp} = A \cos(\theta)$$
  
$$\Phi = vA_{\perp} = vA \cos(\theta) = \vec{v} \cdot \vec{A}$$

# GAUSS' LAW FOR THE GRAVITATIONAL FIELD

In our earliest work with gravitational force problems, such as calculating the motion of the moon or artificial earth satellites, we got the correct answer by replacing the extended spherical earth by a point mass  $M_e$  located at the center of the earth. It is surprising that the gravitational force exerted on you by every rock, mountain, body of water, the earth's iron core, etc. all adds up to be equivalent to the force that would be exerted by a point mass  $M_e$  located 6,000 km beneath you. A simplified version of history is that Isaac Newton delayed his publication of the theory of gravity 20 years, and invented calculus, in order to show that the gravitational force of the entire earth was equivalent to the force exerted by a point mass located at the center.

One can do a brute force calculus derivation to prove the above result, or one can get the result almost immediately from Gauss'law. With Gauss'law, we can also find out, almost by inspection, how the gravitational force decreases as we go down inside the earth.

Since gravity and electricity are both  $1/r^2$  forces, Gauss' law also applies to gravity, and we can get the formulas for the gravitational version by comparing the constants that appear in the force laws. Defining the gravitational field  $\vec{g}$  as the force on a unit mass (note that this is also the acceleration due to gravity), we have for a point charge and a point mass

$$electric \vec{E} = \frac{Q}{4\pi\epsilon_{o}r^{2}} \qquad (48)$$

$$gravitational \vec{g} = \frac{GM}{r^{2}} \qquad (49)$$

Aside from the fact that gravitational forces are always attractive (therefore the field lines always go into a mass m) the only other difference is that  $1/4\pi\epsilon_0$  is replaced by G

$$G \Leftrightarrow \frac{1}{4\pi\varepsilon_{o}} \tag{50}$$

For the electric forces, a point charge Q produces an amount of flux

$$\Phi_{\rm E} = \frac{Q}{\varepsilon_{\rm o}} = \frac{1}{4\pi \varepsilon_{\rm o}} \times 4\pi \, Q \tag{51}$$

Replacing  $1/4\pi \epsilon_o$  by G and Q by M, we expect that a mass M destroys (rather than creates) an amount of flux  $\Phi_G$  given by

$$\Phi_{\rm G} = {\rm G}(4\pi{\rm M}) = 4\pi{\rm G}{\rm M} \tag{52}$$

## **Gravitational Field of a Point Mass**

Let us check if we have the correct flux formula by first calculating the gravitational field of a point mass. In Figure (30) we have a point mass M surrounded by a Gaussian spherical surface of radius r. The flux flowing into the point mass is given by Equation (52) as  $4\pi$ GM. The flux flowing in through the Gaussian surface is

$$\Phi_{G}\begin{pmatrix}flux in\\through\\Gaussian\\surface\end{pmatrix} = \vec{g} \cdot \vec{A} = gA_{\perp} = g(4\pi r^{2})$$
(53)

where the area of the sphere is  $4\pi r^2$ . Equating the flux in through the sphere (53) to the flux into M (52) gives

$$g(4\pi r^2) = 4\pi GM$$

$$g = \frac{GM}{r^2}$$
(54)

which is the gravitational force exerted on a unit mass.



**Figure 30** Calculating the gravitational field  $\bar{g}$ of a point mass using Gauss' law.

## Gravitational Field of a Spherical Mass

Now let us model the earth as a uniform sphere of mass as shown in Figure (31). By symmetry the gravitational field lines must flow radially inward toward the center of the sphere. If we draw a Gaussian surface of radius r outside the earth, we have a total flux flowing in through the sphere given by

$$\Phi_{\rm in} = \vec{g} \cdot \vec{A} = gA_{\perp}$$
  
=  $g4\pi r^2$  (55)

Now the total amount of mass inside the sphere is  $M_e$ , so that the total amount of flux that must stop somewhere inside the Gaussian surface is

$$\Phi_{\rm G} = 4\pi \rm{GM}_{\rm e} \tag{56}$$

Since Equations (56) and (55) are identical to Equations (52) and (53) for a point mass, we must get the same answer. Therefore the gravitational field outside a spherically symmetric mass  $M_e$  is the same as the field of a point mass  $M_e$  located at the center of the sphere.



#### Figure 31

Since  $4\pi GM_e$  lines of flux go in through the mapping surface in both Figures 30 and 31, the field at the mapping surface must be the same for both. This is why, when we are above the surface of the earth, we can treat the earth as a point mass located at the center of the earth.

## **Gravitational Field Inside the Earth**

The gravitational field outside a spherical mass is so easy to calculate using Gauss' law, that you might suspect that we really haven't done anything. The calculation becomes more interesting when we go down inside the earth and the field is no longer that of a point mass  $M_e$ . By determining how the field decreases as we go inside the earth, we can gain some confidence in the calculational capabilities of Gauss' law.

In Figure (32), we are representing the earth by a uniform sphere of matter of mass  $M_e$  and radius  $R_e$ . Inside the earth we have drawn a Gaussian surface of radius r, and assume that the gravitational field has a strength g(r) at this radius. Thus the total flux in through the Gaussian surface  $\Phi_{in}$  is given by

$$\Phi_{\rm in} = \vec{g}(\mathbf{r}) \cdot \vec{A} = g(\mathbf{r}) A_{\perp}$$
  
= g(r) 4\pi r^2 (57)

Now the amount of mass inside our Gaussian surface is no longer  $M_e$ , but only the fraction of  $M_e$  lying below the radius r. That amount is equal to  $M_e$  times the ratio



#### Figure 32

To calculate the gravitational field inside the earth, we draw a mapping surface inside, at a radius r less than the earth radius  $R_e$ . The amount of mass  $M_{in}$ inside the mapping surface is equal to the earth mass  $M_e$  times the ratio of the volume of the mapping sphere to the volume of the earth. of the volume of a sphere of radius r to the volume of the entire earth.

$$M_{r}\left(\frac{mass}{belowr}\right) = M_{e} \frac{\frac{4}{3}\pi r^{3}}{\frac{4}{3}\pi R_{e}^{3}}$$
$$= M_{e} \frac{r^{3}}{R_{e}^{3}}$$
(58)

The amount of flux that the mass  $M_r$  absorbs is given by Equation (52) as

$$\Phi_{\rm G} = 4\pi {\rm GM}_{\rm r} = 4\pi {\rm GM}_{\rm e} \frac{r^3}{{\rm R}_{\rm e}^{\ 3}}$$
(59)

Equating the flux flowing in through the Gaussian surface (Equation (57) to the flux absorbed by  $M_r$  (Equation 59) gives

 $g(r)4\pi r^2 = 4\pi GM_e \frac{r^3}{R_e^3}$ 

or

$$g(r) = GM_e \frac{r}{R^3}$$
 Gravitational field (60)

This result, which can be obtained by a much more difficult calculus calculation, shows the earth's gravitational field dropping linearly (proportional to r), going to zero as r goes to 0 at the center of the earth.

Figure (32) gives an even more general picture of how the earth's gravitational field changes as we go down inside. The flux going in past our Gaussian surface is determined entirely by the mass inside the surface. *The mass in the spherical shell outside the Gaussian surface has no effect at all!* If we are down inside the earth, a distance  $R_i$  from the center, we can accurately determine the gravitational force on us by assuming that all the mass below us (r <  $R_i$ ) is located at a point at the center of the earth, and all the mass above (r >  $R_i$ ) does not exist.

#### **Exercise 8**

As shown in Figure (33) a plastic ball of radius R has a total charge Q uniformly distributed throughout it. Use Gauss' law to:

a) Calculate the electric field  $\vec{E}(r)$  outside the sphere (r > R). How does this compare with the electric field of a point charge?

b) Calculate the electric field inside the plastic sphere (r < R).

(Try to do this now. The solution is on the next page as an example.)



#### Figure 33

Diagram for exercises 8&10. A plastic sphere of radius R has a charge Q spread uniformly throughout. The problem is to calculate the electric field inside and outside.

# **Solving Gauss' Law Problems**

Using Gauss' law to solve for electric fields can be handled in a relatively straightforward way using the following steps:

- 1) Carefully sketch the problem.
- 2) Draw a mapping surface that passes through the point where you want to solve for the field. Construct the surface so that any field lines going through the surface are perpendicular to the surface. This way you can immediately spot the perpendicular area  $A_{\perp}$ .
- 3) Identify Q<sub>in</sub>, the amount of electric charge inside your mapping surface.
- 4) Solve for E using the Equation  $\Phi = EA_{\perp} = Q_{in}/\epsilon_o$ .
- 5) Check that your answer is reasonable.

As an example, let us follow these steps to solve part (b) of Exercise 8, i.e., find the electric field E inside a uniform ball of charge.

 Sketch the problem. The sphere has a radius R, and total charge +Q. By symmetry the electric field must go radially outward for a positive charge (or radially inward for a negative charge).



2) Since we want the field inside the charged sphere, we will use a spherical mapping surface of radius r < R. Because the electric field is everywhere perpendicular to the mapping surface, the area of the mapping surface is  $A_{\perp} = 4\pi r^2$ .

3) The simplest way to calculate the amount of charge  $Q_{in}$  inside our mapping surface is to note that since the charge is uniformly spread throughout the sphere,  $Q_{in}$  is equal to the total charge Q times the ratio of the volume inside the mapping surface to the total volume of the charged sphere; i.e.,

$$Q_{in} = Q \times \frac{(4/3)\pi r^3}{(4/3)\pi R^3} = \frac{Qr^3}{R^3}$$

4) Now use Gauss' law to calculate E:

$$\Phi = EA_{\perp} = \frac{Q_{in}}{\varepsilon_o}$$
$$E \times 4\pi r^2 = \frac{Qr^3}{\varepsilon_o R^3}$$

E	=	Qr
		$4\pi\epsilon_{o}R^{3}$

5) Check to see if the answer is reasonable. At r = 0, we get E = 0. That is good, because at the center of the sphere, there is no unique direction for  $\vec{E}$  to point. At r = R, our formula for E reduces to  $E = Q/(4\pi\epsilon_0 R^2)$ , which is the field of a point charge Q when we are a distance R away. This agrees with the idea that once we are outside a spherical charge, the electric field is the same as if all the charge were at a point at the center of the sphere.

## Exercise 9

As shown in Figure (34), the inside of the plastic sphere has been hollowed out. The total charge on the sphere is Q. Use Gauss' law to

*a)* Determine the strength of the electric field inside the hollow cavity.

*b)* Calculate the strength of the electric field inside the plastic.

*c*) Calculate the strength of the electric field outside the plastic.



## Figure 34

Diagram for Exercises 9 & 11. The charge is now spread throughout a spherical shell.

## Exercise 10

Repeat *Exercise 8*, assuming that Figure (33) represents the end view of a very long charged plastic rod with a charge of  $\lambda$  coulombs per meter. (A section of length L will thus have a charge Q =  $\lambda$  L.)

## Exercise 11

Repeat *Exercise 9*, assuming that Figure (34) represents the end view of a charged hollow plastic rod with a charge of  $\lambda$  coulombs per meter.

## Exercise 12

A hydrogen atom (H atom) consists of a proton with an electron moving about it. The classical picture is that the electron orbits about the proton much like the earth orbits the sun. A model that has its origins in quantum mechanics and is more useful to chemists, is to picture the electron as being smeared out, forming a ball of negative charge surrounding the proton. This ball of negative charge is called an "electron cloud ".

For this problem, assume that the electron cloud is a *uniform* sphere of negative charge, a sphere of radius R centered on the proton as shown in Figure (35). The total negative charge (–e) in the electron cloud just balances the positive charge (+e) on the proton, so that the net charge on the H atom is zero.

*a)* Sketch the electric field for this model of the H atom. Show the electric field both inside and outside the electron cloud.

**b)** Calculate the magnitude of the electric field for both r < R (inside the cloud) and r > R (outside the cloud).



## Figure 35

Picture a hydrogen atom as a proton (of charge +e), surrounded by an electron cloud. Think of the cloud as a uniform ball of negative charge, with a net charge -e.

## **Exercise 13**

A butterfly net with a circular opening of radius R, is in a uniform electric field of magnitude E as shown in Figure (36). The opening is perpendicular to the field. Calculate the net flux of the electric field through the net itself. (The amount of flux through each hole in the net is  $\vec{E} \cdot d\vec{A}$  where  $d\vec{A}$  is the area of the hole.) (This is one of our favorite problems from Halliday and Resnick.)



#### Figure 36

Butterfly net in a uniform electric field. With Gauss' law, you can easily calculate the flux of the electric field through the net itself.

## Exercise 14

Electric fields exist in the earth's atmosphere. (You get lightning if they get too strong.)

On a particular day, it is observed that at an altitude of 300 meters, there is a downward directed electric field of magnitude

$$E(300 \text{ m}) = 70 \frac{\text{newton}}{\text{coulomb}}$$

Down at an altitude of 200 meters, the electric field still points down, but the magnitude has increased to

$$E(200 m) = 100 \frac{newton}{coulomb}$$

How much electric charge is contained in a cube 100 meters on a side in this region of the atmosphere?



**Figure 37** *Electric field at two different altitudes.* 

# **PROBLEM SOLVING**

One can devise a number of Gauss' law problems where one plugs various numerical values into the formulas we have derived. But that is not the point of this chapter. Here we are interested, not so much in the answers, as in the concepts and techniques used to derive them.

In this chapter we have introduced two new concepts. One is electric flux  $\Phi = EA_{\perp}$ , and the other is that the total flux out through a closed surface is equal to  $Q_{inside}/\epsilon_0$ . These two concepts allow us to easily solve for the electric field in certain special cases. Those cases are where  $A_{\perp}$  is a sphere, a cylinder, or a plane.

What you need to get from this discussion is the beginning of an intuitive picture of electric flux, how it is related to the flow of an incompressible fluid, and how this concept can be used to handle the few but important examples where  $A_{\perp}$  is either a sphere, cylinder, or plane. Numerical applications can come later, now is the time to develop intuition.

Most students have some difficulty handling Gauss's law problems the first time they see them because the concepts involved are new and unfamiliar. Then when the problems are solved in a homework session, a common reaction is, "Oh, those are not so hard after all". One gets the feeling that by just watching the problem solved, and seeing that it is fairly easy after all, they understand it. The rude shock comes at an exam where suddenly the problem that looked so easy, has become unsolvable again.

There is a way to study to avoid this rude shock. Pick one of the problems you could not solve on your own, a problem you saw solved in class or on an answer sheet. A problem that looked so easy after you saw it solved.

Wait a day or two after you saw the solution, clean off your desk, take out a blank sheet of paper, and try solving the problem.

Something awful may happen. That problem that looked so easy in the homework review session is now impossible again. You can't see how to do it, and it looked so easy two days ago. You feel really bad—but don't, it happens to everyone. Instead, if you cannot get it, just peek at the solution to see what point you missed, then put the solution away and solve it on your own. You may have to peek a couple of times, but that is OK.

If you had to peek at the solution, then wait another day or so, clean off your desk, and try again. Soon you will get the solution without looking, and you will not forget how to solve that problem. You will get more out of this technique than solving 15 numerical examples.

When you are studying a new topic with new, unfamiliar concepts, the best way to learn the subject is to thoroughly learn a few, well chosen worked out examples. By learn, we mean problems you can work on a blank sheet of paper without looking at a solution.

Pick examples that are relatively simple but clearly illustrate the concepts involved. For this chapter, one could pick the example of calculating the electric field inside and outside a uniform ball of charge. If you can do that problem on a clean desk, you can probably do most of the other problems in this chapter without too much difficulty.

Why learn a sample problem for each new topic? The reason is that if you know one worked example you will find it easy to remember the entire topic. That worked example reminds you immediately how that concept works, how it functions.

In this text, Chapters 24-32 on electric and magnetic fields involve many new concepts. Concepts you will not have seen unless you have already taken the course. As we go along, we will suggest sample problems, what we call "clean desk problems", which serve as a good example of the way the new concept is used. You may wish to choose different sample problems, but the best way to learn this topic is to develop a repertory of selected sample problems you understand cold.

At this point, go back to some of the problems in this chapter, particularly Exercises 9 through 14 and see if you can solve them on a clean desk. If you can, you are ready for the next chapter.

# Chapter 25 Field Plots and Electric Potential

Calculating the electric field of any but the simplest distribution of charges can be a challenging task. Gauss' law works well where there is considerable symmetry, as in the case of spheres or infinite lines of charge. At the beginning of Chapter (24), we were able to use a brute force calculus calculation to determine the electric field of a short charged rod. But to handle more complex charge distributions we will find it helpful to apply the techniques developed by map makers to describe complex terrains on aflat map. This is the technique of the contour map which works equally well for mapping electric fields and mountain ranges. Using the contour map ideas, we will be lead to the concept of a **potential** and **equipotential lines** or **surfaces**, which is the main topic of this chapter.

## THE CONTOUR MAP

Figure (1) is a contour map of a small island. The contour lines, labeled 0, 10, 20, 30 and 40 are lines of equal height. Anywhere along the line marked 10 the land is 10 meters above sea level. (You have to look at some note on the map that tells you that height is measured in meters, rather than feet or yards.)

You can get a reasonable understanding of the terrain just by looking at the contour lines. On the south side of the island where the contour lines are far apart, the land slopes gradually upward. This is probably where the beach is located. On the north side where the contour lines are close together, the land drops off sharply. We would expect to see a cliff on this side of the island.



#### Figure 1

Contour map of a small island with a beach on the south shore, two hills, and a cliff on the northwest side. The slope of the island is gradual where the lines are far apart, and steep where the lines are close together. If you were standing at the point labeled (A) and the surface were slippery, you would start to slide in the direction of the arrow. Although we would rather picture this island as being in the south seas, imagine that it is in the North Atlantic and a storm has just covered it with a sheet of ice. You are standing at the point labeled A in Figure (1), and start to slip. If the surface is smooth, which way would you start to slip?

A contour line runs through Point A which we have shown in an enlargement in Figure (2). You would not start to slide along the contour line because all the points along the contour line are at the same height. Instead, you would start to slide in the steepest downhill direction, which is perpendicular to the contour line as shown by the arrow.

If you do not believe that the direction of steepest descent is perpendicular to the contour line, choose any smooth surface like the top of a rock, mark a horizontal line (an equal height line) for a contour line, and carefully look for the directions that are most steeply sloped down. You will see that all along the contour line the steepest slope is, in fact, perpendicular to the contour line.



## Figure 2

Along a contour line the land is level. The direction of steepest slope or descent is perpendicular to the contour line. Skiers are familiar with this concept. When you want to stop and rest and the slope is icy, you plant your skis along a contour line so that they will not slide either forward or backward. The direction of steepest descent is now perpendicular to your skis, in a direction that ski instructors call the *fall line*. The fall line is the direction you will start to slide if the edges of your skis fail to hold.

In Figure (3), we have redrawn our contour map of the island, but have added a set of perpendicular lines to show the directions of steepest descent, the direction of the net force on you if you were sitting on a slippery surface. These lines of steepest descent, are also called *lines of force*. They can be sketched by hand, using the rule that the lines of force must always be perpendicular to the contour lines.



#### Figure 3

You can sketch in the lines of steepest descent by drawing a set of lines that are always perpendicular to the contour lines. These lines indicate the direction a ball would <u>start</u> to roll if placed at a point on the line. In Figure (4) we have the same island, but except for the zero height contour outlining the island, we show only the lines of force. The exercise here, which you should do now, is sketch in the contour lines. Just use the rule that the contour lines must be drawn perpendicular to the lines of force. The point is that you can go either way. Given the contour lines you can sketch the lines of force, or given the lines of force you can sketch the contour lines. This turns out to be a powerful technique in the mapping of any complex physical or mathematical terrain.



#### Figure 4

Here we have removed the contour lines leaving only the lines of steepest descent and the outline of the island. By drawing a set of lines perpendicular to the lines of steepest descent, you can more or less reconstruct the contour lines. The idea is that you can go back and forth from one set of lines to the other.

## EQUIPOTENTIAL LINES

On a contour map of an island, the contour lines are lines of equal height. If you walk along a contour line, your height h, and therefore your gravitational potential energy mgh, remains constant. As a result, we can call these the lines of constant or equal potential energy, *equipotential* energy lines for short.

Let us apply these mapping concepts to the simpler situation of a spherical mass M shown in Figure (5). As in Chapter 24, we have drawn the gravitational field lines, which point radially inward toward the center of M. We determined these field lines by placing a test particle of mass m in the vicinity of M as shown in Figure (6). The potential energy of this test mass m is given by our old formula (see Equation 10-50a)

GMm	attractive forces	
$PE = -\frac{GRM}{r}$	havenegative	(1)
I	potentialenergy	

In our discussion of electric and gravitational fields, we defined the fields as the force on a unit test particle. Setting m = 1 for a unit test mass, we get as the formula for the potential energy of our unit test mass

Potential energy of  
a unit test mass 
$$= -\frac{GM}{r}$$
 (2)

equipotential

lines



**Figure 6** The gravitational potential energy of two masses separated by a distance r is -GMm/r.

#### Figure 5

The gravitational field lines for a spherical mass point radially inward. The lines of constant potential energy are circles of equal height above the mass. The equipotential lines are everywhere perpendicular to the field lines, just as, in the map of the island, the contour lines were everywhere perpendicular to the lines of steepest descent. From Equation (2) we see that if we stay a constant distance r out from M, if we are one of the concentric circles in Figure (5), then the potential energy of the unit test mass remains constant. These circles, drawn perpendicular to the lines of force, are again equal potential energy lines.

There is a convention in physics to use the word *potential* when talking about the potential energy of a unit mass or unit charge. With this convention, then - GM/r in Equation (2) is the formula for the gravitational potential of a mass M, and the constant radius circles in Figure (5) are lines of constant potential. Thus the name *equipotential lines* for these circles is fitting.

# Negative and Positive Potential Energy

In Figure (7) we have drawn the electric field lines of a point charge Q and drawn the set of concentric circles perpendicular to the field lines as shown. From the close analogy between the electric and gravitational force, we expect that these circles represent lines of constant electric potential energy, that they are the electric equipotential lines.



## Figure 7

The <u>electric potential</u> is the potential energy of a positive unit test charge  $q_{test} = +1$  coulomb. Because a positive charge + Q and a positive test charge repel this potential energy is positive. But there is one important difference between Figures (5) and (7). In Figure (5) the gravitational force on our unit test mass is attractive, in toward the mass M. In Figure (7), the force on our unit positive test particle is out, away from Q if Q is positive. When we have an attractive force as in Figure (5), the potential energy is negative as in Equation (1). But when the force is repulsive, as in Figure (7), the potential energy is positive. Let us briefly review the physical origin for this difference in the sign of the potential energy.

In any discussion of potential energy, it is necessary to define the zero of potential energy, i.e. to say where the floor is. In the case of satellite motion, we defined the satellite's potential energy as being zero when the satellite was infinitely far away from the planet. If we release a satellite at rest a great distance from the planet, it will start falling toward the planet. As it falls, it gains kinetic energy, which it must get at the expense of gravitational potential energy. Since the satellite started with zero gravitational potential energy when far out and loses potential energy as it falls in, it must end up with negative potential energy when it is near the planet. This is the physical origin of the minus sign in Equation (1). Using the convention that potential energy is zero at infinity, then attractive forces lead to negative potential energies.

If the force is repulsive as in Figure (7), then we have to do work on our test particle in order to bring it in from infinity. The work we do against the repulsive force is stored up as positive potential energy which could be released if we let go of the test particle (and the test particle goes flying out). Thus the convention that potential energy is zero at infinity leads to positive potential energies for repulsive forces like that shown in Figure (7).

# ELECTRIC POTENTIAL OF A POINT CHARGE

Using the fact that we can go from the gravitational force law to Coulomb's law by replacing GMm by  $Qq/4\pi\epsilon_0$  (see Exercise 1), we expect that the formula for the electric potential energy of a charge q a distance r from Q is

electric potential energy  
of a charge q 
$$\left| \right\rangle = + \frac{Qq}{4\pi\epsilon_0 r}$$
 (3)

The + sign in Equation (3) indicates that for positive Q and q we have a repulsive force and positive potential energy. (If Q is negative, but q still positive, the force is attractive and the potential energy must be negative.)

To determine the potential energy of a *unit test charge*, we set q = 1 in Equation (3) to get

electric potential  
energy of a unit  
test charge 
$$= \begin{pmatrix} electric \\ potential \end{pmatrix} = \frac{Qq}{4\pi\epsilon_0 r}$$

(4)

Following the same convention we used for gravity, we will use the name *electric potential* for the potential energy of a unit test charge. Thus Equation (4) is the formula for the electric potential in the region surrounding the charge Q. As expected, the lines of *equal potential*, the *equipotential lines* are the circles of constant radius seen in Figure (7).

#### **Exercise 1**

Start with Newton's gravitational force law, replace GMm by  $Qq/4\pi\epsilon_0$ , and show that you end up with Coulomb's electrical force law.

## **CONSERVATIVE FORCES**

(This is a formal aside to introduce a point that we will treat in much more detail later.)

Suppose we have a fixed charge Q and a small test particle q as shown in Figure (8). The potential energy of q is defined as zero when it is infinitely far away from Q. If we carry q in from infinity to a distance r, we do an amount of work on the particle

Work we do 
$$\Big\} = \int_{\infty}^{r} \vec{F}_{us} \cdot d\vec{x}$$
 (5)

If we apply just enough force to overcome the electric repulsive force, if  $\vec{F}_{us} = -q\vec{E}$ , then the work we do should all be stored as electric potential energy, and Equation (5), with  $\vec{F}_{us} = -q\vec{E}$  should give us the correct electric potential energy of the charge q.

But an interesting question arises. Suppose we bring the charge q in along two different paths, paths (1) and (2) shown in Figure (8). Do we do the same amount of work, store the same potential energy for the two different paths?



#### Figure 8

If we bring a test particle q in from infinity to a distance r from the charge Q, the electric potential energy equals  $Qq/4\pi\varepsilon_o r$ . But this potential energy is the work we do in bringing q in from infinity:

$$\int_{\infty}^{r} \vec{F}_{us} \cdot d\vec{r} = \frac{Qq}{4\pi\varepsilon_{o}r}$$

This answer does not depend upon the path we take bringing q in.

If we lift an eraser off the floor up to a height h, and hold it still, then it does not matter what path we took, the net amount of work we did was mgh and this is stored as gravitational potential energy. When the work we do against a force depends only on the initial and final points, and not on the path we take, we say that the force is conservative.

In contrast, if we move the eraser over a horizontal table from one point to another, the amount of work we do against friction depends very much on the path. The longer the path the more work we do. As a result we cannot define a **friction potential energy** because it has no unique value. Friction is a *non-conservative force*, and non-conservative forces do not have unique potential energies.

The gravitational fields of stationary masses and the electric fields of stationary charges all produce conservative forces, and therefore have unique potential energies. We will see however that moving charges can produce electric fields that are not conservative! When that happens, we will have to take a very careful look at our picture of electric potential energy. But in dealing with the electric fields of static charges, as we will for a few chapters, we will have unique electric potential energies, and maps of equipotential lines will have an unambiguous meaning.

# ELECTRIC VOLTAGE

In our discussion of Bernoulli's equation, we gave the collection of terms (P +  $\rho$ gh +  $1/2\rho$ v<sup>2</sup>) the name *hydrodynamic voltage*. The content of Bernoulli's equation is that this hydrodynamic voltage is constant along a stream line when the fluid is incompressible and viscous forces can be neglected. Two of the three terms,  $\rho$ gh and  $1/2\rho$ v<sup>2</sup> represent the energy of a unit volume of the fluid, thus we see that our hydrodynamic voltage has the dimensions of energy per unit volume.

*Electric voltage* is a quantity with the dimensions of *energy per unit charge* that in different situations is represented by a series of terms like the terms in Bernoulli's hydrodynamic voltage. There is the potential energy of an electric field, the chemical energy supplied by a battery, even a kinetic energy term, seen in careful studies of superconductors, that is strictly analogous to the  $1/2\rho v^2$  term in Bernoulli's equation. In other words, electric voltage is a complex concept, but it has one simplifying feature. Electric voltages are measured by a common experimental device called a voltmeter. In fact we will take as the definition of electric voltage, that quantity which we measure using a voltmeter.

This sounds like a nebulous definition. Without telling you how a voltmeter works, how are you to know what the meter is measuring? To overcome this objection, we will build up our understanding of what a voltmeter measures by considering the various possible sources of voltage one at a time. Bernoulli's equation gave us all the hydrodynamic voltage terms at once. For electric voltage we will have to dig them out as we find them.

Our first example of an electric voltage term is the *electric potential energy of a unit test charge*. This has the dimensions of energy per unit charge which in the MKS system is joules/coulomb and called volts.

$$1 \frac{\text{joule}}{\text{Coulomb}} \equiv 1 \text{ volt}$$
(6)

In Figure (9), which is a repeat of Figure (8) showing the electric field lines and equipotential lines for a point charge Q, we see from Equation (4) that a unit test particle at Point (1) has a potential energy, or voltage  $V_1$ given by

$$V_{1} = \frac{Q}{4\pi\epsilon_{0}r_{1}} \qquad electric potential or voltage at Point(1)$$

At Point (2), the electric potential or voltage  $V_2$  is given by

$$V_2 = \frac{Q}{4\pi\epsilon_0 r_2} \qquad electric potential or voltage at Point(2)$$

Voltmeters have the property that they only measure the *difference in voltage* between two points. Thus if we put one lead of a voltmeter at Point (1), and the other at Point (2) as shown, then we get a voltage reading V given by

voltmeter  
reading 
$$V \equiv V_2 - V_1 = \frac{Q}{4\pi\epsilon_0} \left(\frac{1}{r_2} - \frac{1}{r_1}\right)$$

If we put the two voltmeter leads at points equal distances from Q, i.e. if  $r_1 = r_2$ , then the voltmeter would read zero. Since the voltage difference between any two points on an equipotential line is zero, the voltmeter reading must also be zero when the leads are attached to any two points on an equipotential line.



Figure 9

A voltmeter measures the difference in electrical voltage between two points.

This observation suggests an experimental way to map equipotential lines or surfaces. Attach one lead of the voltmeter to some particular point, call it Point (A). Then move the other lead around. Whenever you get a zero reading on the voltmeter, the second lead must be at another point of the same equipotential line as Point (A). By marking all the points where the meter reads zero, you get a picture of the equipotential line.

The discussion we have just given for finding the equipotential lines surrounding a point charge Q is not practical. This involves electrostatic measurements that are extremely difficult to carry out. Just the damp air from your breath would affect the voltages surrounding a point charge, and typical voltmeters found in the lab cannot make electrostatic measurements. Sophisticated meters in carefully controlled environments are required for this work.

But the idea of potential plotting can be illustrated nicely by the simple laboratory apparatus illustrated in Figure (10). In that apparatus we have a tray of water (slightly salty or dirty, so that it is somewhat conductive), and two metal cylinders attached by wire leads to a battery as shown. There are also two probes consisting of a bent, stiff wire attached to a block of wood and adjusted so that the tips of the wires stick down in the water. The other end of the probes are attached to a voltmeter so we can read the voltage difference between the two points (A) and (B), where the probes touch the water.



## Figure 10

Simple setup for plotting fields. You plot equipotentials by placing one probe (A) at a given position and moving the other (B) around. Whenever the voltage V on the voltmeter reads zero, the probes are at points of equipotential.



#### Figure 11

Plot of the equipotential lines from a student project by B. J. Grattan. Instead of a tray of water, Grattan used a sheet of conductive paper, painting two circles with aluminum paint to replace the brass cylinders. (The conductive paper and the tray of water give similar results.) We used the Adobe Illustrator program to draw the lines through Grattan's data points. If we keep Probe (A) fixed and move Probe (B) around, whenever the voltmeter reads zero, Probe (B) will be on the equipotential line that goes through Point (A). Without too much effort, one can get a complete plot of the equipotential line. Each time we move Probe (A) we can plot a new equipotential line. A plot of a series of equipotential lines is shown in Figure (11).

Once we have the equipotential lines shown in Figure (11), we can sketch the lines of force by drawing a set of lines perpendicular to the equipotential as we did in Figure (12). With a little practice you can sketch fairly accurate plots, and the beauty of the process is that you did not have to do any calculations!



#### Figure 12

It does not take too much practice to sketch in the field lines. Draw smooth lines, always perpendicular to the equipotential lines, and maintain any symmetry that should be there.

#### **Exercise 2**

The equipotential plot of Figure (11) and the field lines of Figure (12) were taken from a student project. The field lines look like the field of two point charges +Q and -Q separated by a distance r. But who knows what is happening in the shallow tank of water (or a sheet of conducting paper)? Perhaps the field lines more nearly represent the field of two line charges + $\lambda$  and - $\lambda$ separated by a distance r.

The field of a point charge drops off as  $1/r^2$  while the field of a line charge drops off as 1/r. The point of the exercise is to decide whether the field lines in Figure (12) (or your own field plot if you have constructed one in the lab) more closely represent the field of a point or a line charge.



Hint—Look at the electric field at Point A in Figure (12), enlarged in Figure (13). We know that the field E at Point A is made up of two components,  $\vec{E}_1$  directed away from the left hand cylinder, and  $\vec{E}_2$  directed toward the right hand cylinder, and the net field E is the vector sum of the two components. If the field is the field of point charges then  $E_1$  drops off as  $1/r_1^2$  and  $E_2$  as  $1/r_2^2$ . But if the field is that of line charges,  $|\vec{E}_1|$  drops off as  $1/r_1$  and  $|\vec{E}_2|$  as  $1/r_2$ . We have chosen Point (A) so that  $r_1$ , the distance from (A) to the left cylinder is quite a bit longer than the distance r<sub>2</sub> to the right cylinder. As a result, the ratio of  $|\vec{E}_1|$  to  $|\vec{E}_2|$  and thus the direction of  $\vec{E}$ , will be quite different for 1/r and 1/r<sup>2</sup> forces. This difference is great enough that you can decide, even from student lab results, whether you are looking at the field of point or line charges. Try it yourself and see which way it comes out.

#### Figure 13

Knowing the direction of the electric field at Point (A) allows us to determine the relative magnitude of the fields  $\vec{E}_1$  and  $\vec{E}_2$  produced by charges 1 and 2 alone. At Point (A), construct a vector  $\vec{E}$  of convenient length parallel to the field line through (A). Then decompose  $\vec{E}$  into component vectors  $\vec{E}_1$  and  $\vec{E}_2$ , where  $\vec{E}_1$  lies along the line from charge 1 to Point (A), and  $\vec{E}_2$ along the line toward charge 2. Then adjust the lengths of  $\vec{E}_1$  and  $\vec{E}_2$  so that their vector sum is  $\vec{E}$ .



#### Figure 14

Model of the electric field in the region of two point charges  $Q_+ = +3$ ,  $Q_- = -1$ . Using the analogy to a topographical map, we cut out plywood slabs in the shape of the equipotentials from the computer plot of Figure 15, and stacked the slabs to form a three dimensional surface. The field lines, which are marked with narrow black tape on the model, always lead in the direction of steepest descent on the surface.



## **A Field Plot Model**

The analogy between a field plot and a map maker's contour plot can be made even more obvious by constructing a plywood model like that shown in Figure (14).

To construct the model, we made a computer plot of the electric field of charge distribution consisting of a charge +3 and -1 seen in Figure (15). We enlarged the computer plot and then cut out pieces of plywood that had the shapes of the contour lines. The pieces of plywood were stacked on top of each other and glued together to produce the three dimensional view of the field structure.

In this model, each additional thickness of plywood represents one more equal step in the electric potential or voltage. The voltage of the positive charge Q = +3 is represented by the fat positive spike that goes up toward  $+\infty$  and the negative charge q = -1 is represented by the smaller hole that heads down to  $-\infty$ . These spikes can be seen in the back view in Figure (14), and the potential plot in Figure (16).

In addition to seeing the contour lines in the slabs of plywood, we have also marked the lines of steepest descent with narrow strips of black tape. These lines of steepest descent are always perpendicular to the contour lines, and are in fact, the electric field lines, when viewed from the top as in the photograph of Figure (15).

Figure (17) is a plywood model of the electric potential for two positive charges, Q = +5, Q = +2. Here we get two hills.



**Figure 17** Model of the electric potential in the region of two point charges Q = +5 and Q = +2.



## Figure 15

Computer plot of the field lines and equipotentials for a charge distribution consisting of a positive charge + 3 and a negative charge - 1. These lines were then used to construct the plywood model.

## **Computer Plots**

There are now many excellent programs that have personal computers draw out field plots for various charge distributions. In most of these programs you enter an array of charges and the computer draws the field and equipotential lines. You should practice with one of these programs in order to develop an intuition for the field structures various charge distributions produce. In particular, try the charge distribution shown in Figure (18) and (19). In Figure (18), we wish to see the field of oppositely charged plates (a positive plate on the left and a negative one on the right). This charge distribution will appear in the next chapter in our discussion of the parallel plate capacitor.

In Figure (19) we are modeling the field of a circle or in 3-dimensions a hollow sphere of charge. Something rather remarkable happens to the electric field lines in this case. Try it and see what happens!

## **Exercise 3**

If you have a computer plotting program available, plot the field lines for the charge distributions shown in Figures (18,19), and explain what the significant features of the plot are.

ł	—
ł	_
ł	_
ł	_
F	_

#### Figure 18

The idea is to use the computer to develop an intuition for the shape of the electric field produced by various distributions of electric charge. Here the parallel lines of charge simulates two plates with opposite charge.



Figure 19

We have placed + charges around a circle to simulate a cylinder or sphere of charge. You get interesting results when you plot the field lines for this distribution of charge.

## **Exercise 4**

Figures (20a) and (20b) are computer plots of the electric field of opposite charges. One of the plots represents the  $1/r^2$  field of 3 dimensional point charges. The other is the end view of the 1/r field of line charges. You are to decide which is which, explaining how you can tell.



#### Figure 20

Computer plots of 1/r and  $1/r^2$  (two dimensional and three dimensional) fields of equal and opposite charges. You are to figure out which is which.

# Chapter 26 Electric Fields and Conductors

In this chapter we will first discuss the behavior of electric fields in the presence of conductors, and then apply the results to three practical devices, the Van de Graaff generator, the electron gun, and the parallel plate capacitor. Each of these examples provides not only an explanation of a practical device, but also helps build an intuitive picture of the concept of electric voltage.

# ELECTRIC FIELD INSIDE A CONDUCTOR

If we have a piece of metal a few centimeters across as illustrated in Figure (1), and suddenly turn on an electric field, what happens? Initially the field goes right through the metal. But within a few pico seconds (1 pico second =  $10^{-12}$  seconds) the electrons in the metal redistribute themselves inside the metal creating their own field that soon cancels the external applied electric field, as indicated in Figure (2).



Figure 1

What is the electric field inside a chunk of metal? Metals have conduction electrons that are free to move. If there were an electric field inside the metal, the conduction electrons would be accelerated by the field.



**Figure 2** If you place a chunk of metal in an external electric field, the electrons move until there is no longer a force on them.

The very concept of an electrical conductor requires that, in the steady state, there be no electric field inside. To see why, imagine that there is a field inside. Since it is a conductor, the electrons in the conductor are free to move. If there is a field inside, the field will exert a force on the electrons and the electrons will move. They will continue to move until there is no force on them, i.e., until there is no field remaining inside. The electrons must continue to move until the field they create just cancels the external field you applied.

## **Surface Charges**

Where does the redistributed charge have to go in order to create an electric field that precisely cancels the applied electric field? Gauss' law provides a remarkably simple answer to this question. The redistributed charge must reside on the *surface of the conductor*. This is because Gauss' law requires that there be no net charge inside the volume of a conductor.

To see why, let us assume that a charge Q is inside a conductor as shown in Figure (3). Draw a small Gaussian surface around Q. Then by Gauss' law the flux  $\Phi = \vec{E} \cdot \vec{A}$  coming out through the Gaussian surface must be equal to  $Q_{in}/\epsilon_0$  where  $Q_{in}$  is the net charge inside the Gaussian surface. But if there is no\_field inside the conductor, if  $\vec{E} = 0$ , then the flux  $\vec{E} \cdot \vec{A}$  out through the Gaussian surface must be zero, and therefore the charge  $Q_{in}$  must be zero.



## Figure 3

Is there any electric charge inside a conductor? To find out, draw a Gaussian surface around the suspected charge. Since there is no electric field inside the conductor, there is no flux out through the surface, and therefore no charge inside. If there is no charge inside the conductor, then the only place any charge can exist is in the surface. If there is a redistribution of charge, the redistributed charge must lie on the surface of the conductor.

Figure (4) is a qualitative sketch of how surface charge can create a field that cancels the applied field.



## Figure 4a





#### Figure 4b

In response to the electric field, the electrons move to the left surface of the metal, leaving behind positive charge on the right surface. These two surface charges have their own field  $\vec{E}'$  that is oppositely directed to  $\vec{E}$ .



#### Figure 4c

Inside the block of metal the fields cancel. The result is that the external field on the left stops on the negative surface charge. The field on the right starts again on the positive surface charge. In Figure (4a) we see the electric field just after it has been turned on. Since the electrons in the metal are negatively charged (q = -e), the force on the electrons  $\vec{F} = (-e)\vec{E}$  is opposite to  $\vec{E}$  and directed to the left.

In Figure (4b), electrons have been sucked over to the left surface of the metal, leaving positive charge on the right surface. The negative charge on the left surface combined with the positive charge on the right produced the left directed field  $\vec{E}'$  shown by the dotted lines. The oppositely directed fields  $\vec{E}$  and  $\vec{E}'$  cancel in Figure (4c) giving no net field inside the metal.

## Surface Charge Density

When a field E impinges on the surface of a conductor, it must be oriented at right angles to the conductor as shown in Figure (5). The reason for this is that if  $\vec{E}$  had a component  $\vec{E}_{||}$  parallel to the surface,  $\vec{E}_{||}$  would pull the movable charge along the surface and change the charge distribution. The only direction the surface of the conductor, thus for a stable setup the electric field at the surface must be perpendicular as shown.



The amount of charge on this surface is  $\sigma dA$ 

#### Figure 5

To calculate the surface charge density, we draw a small cylindrical pill box of cross-sectional area dA. We then equate the flux of electric field out through the right surface of the pill box to  $1/\epsilon_0$  times the charge inside the pill box.

Gauss' law can be used to calculate how much charge must be at the surface if a field of strength  $\vec{E}$  is impinging as shown in Figure (5). In that figure we have drawn a small pill box shaped Gaussian surface, with one end in the conductor and the other outside in the field  $\vec{E}$ . If the area of the end of the pill box is dA, then the flux out of the pill box on the right is  $\Phi_{out} = EdA$ .

Let  $\sigma$  coulombs/meter<sup>2</sup> be the charge density on the surface. The amount of the conductor's surface surrounded by the pill box is dA, thus the amount of charge inside the pill box is

Amount of charge  
inside the  
Gaussian surface 
$$\left. \begin{array}{l} \equiv Q_{in} = \sigma dA \\ \end{array} \right.$$

By Gauss' law, the flux  $\Phi_{out} = EdA$  must equal  $1/\epsilon_0$  times the total charge inside the Gaussian surface and we get

$$\Phi_{\text{out}} = \frac{Q_{\text{in}}}{\varepsilon_0} = \frac{\sigma dA}{\varepsilon_0} = E dA$$

The dA's cancel and we are left with

$$E = \frac{\sigma}{\varepsilon_0} \qquad E = electric field at the conductor \sigma = charge density at the surface (1)$$

Equation (1) gives a simple relation between the strength of the electric field at the surface of a conductor, and the surface charge density  $\sigma$  at that point. Just remember that the field E must be perpendicular to the surface of the conductor. (If the applied field was not originally perpendicular to the surface, surface charges will slide along the surface, reorienting the external field to make it perpendicular.)

To appreciate how far we have come with the concepts of fields and Gauss' law, just imagine trying to derive Equation (1) from Coulomb's law. We wouldn't even know how to begin.

We will now work an example and assign a few exercises to build an intuition for the behavior of fields and conductors. Then we will apply the results to some practical devices.

# Example: Field in a Hollow Metal Sphere

Suppose we have the hollow metal sphere shown in Figure (6). A total charge Q is placed on the sphere. What are the electric fields outside and inside the sphere?

One key to solving this problem is to realize that since the sphere is symmetric, the fields it produces must also be symmetric. We are not interested in fields that do one thing on the left side and something else on the right, for we do not have any physical cause for such an asymmetry.

In Figure (7) we have drawn a Gaussian surface surrounding the metal sphere as shown. Since there is a net charge +Q on the sphere, and therefore inside the Gaussian surface, there must be a net flux  $Q/\epsilon_0$  out through the surface. Since the Gaussian surface has an area  $4\pi r^2$ , Gauss' law gives

$$\Phi = E_{out}A_{\perp} = E_{out} \times 4\pi r^2 = \frac{Q}{\epsilon_0}$$
$$E_{out} = \frac{Q}{4\pi\epsilon_0 r^2}$$
(2)

which happens to be the field of a point charge.

In Figure (8) we have drawn a Gaussian surface inside the metal at a radius  $r_i$ . Since there is no field inside the metal,  $EA_{\perp} = 0$  and there is no flux flowing out through the Gaussian surface. Thus by Gauss' law there can be **no net charge inside the Gaussian surface**. Explicitly this means that there is no surface charge on the inside of the conductor. The charge Q we spread on the conducting sphere **all went to the outside surface**!

Finally in Figure (9) we have drawn a Gaussian surface inside the hollow part of the hollow sphere. Since there is no charge—only empty space inside this Gaussian surface, there can be no flux out through the surface, and the *field*  $\vec{E}$  *inside the hollow part of the sphere is exactly zero*. This is a rather remarkable result considering how little effort was required to obtain it.



#### Figure 6

We place a charge Q on a hollow metal sphere. Where do the charge and the field lines go?





If we place a Gaussian surface around and outside the sphere, we know that the charge Q must be inside the Gaussian surface, and therefore  $Q/\varepsilon_o$ lines must come out through the surface



#### Figure 8

If we place our Gaussian surface inside the metal where  $\vec{E} = 0$ , no lines come out through the Gaussian surface and therefore there must be no net charge Qinside the Gaussian surface. The fact that there is no charge within that surface means all the charge we placed on the sphere spreads to the outside surface.

#### **Exercise 1**

A positive charge +Q is surrounded concentrically by a conducting sphere with an inner radius  $r_a$  and outer radius  $r_b$  as shown in Figure (10). The conducting sphere has no net charge. Using Gauss' law, find the electric field inside the hollow section ( $r < r_a$ ), inside the conducting sphere ( $r_a < r < r_b$ ) and outside the sphere ( $r > r_b$ ). Also calculate the surface charge densities on the inner and outer surfaces of the conducting sphere. Show that Equation (1) applies to the charge densities you calculate.



#### Figure 10

Start with an uncharged hollow metal sphere and place a charge +Q inside. Use Gauss' law to determine the electric field and the surface charges throughout the region.



#### Figure 9

If the Gaussian surface is drawn inside the hollow cavity as shown, then there is no charge inside the Gaussian surface. Thus no field lines emerge through the Gaussian surface, and  $\vec{E}$  must be zero inside the cavity.

#### **Exercise 2**

A chunk of metal has an irregularly shaped cavity inside as shown in Figure (11). There are no holes and the cavity is completely surrounded by metal.

The metal chunk is struck by lightning which produces huge electric fields and deposits an unknown amount of charge on the metal, but does not burn a hole into the cavity. Show that the lightning does not create an electric field inside the cavity. (For a time on the order of pico seconds, an electric field will penetrate into the metal, but if the metal is a good conductor like silver or copper, the distance will be very short.)

(What does this problem have to do with the advice to stay in a car during a thundershower?)



#### Figure 11

A chunk of metal with a completely enclosed hollow cavity inside is struck by lightning.

#### **Exercise 3**

A positive charge +Q placed on a conducting sphere of radius R, produces the electric field shown.

a) What is the charge density  $\sigma$  on the surface of the sphere?

nsity ere?

R

b) Use Equation (1) to find a formula for the magnitude of the electric field  $\vec{E}$  produced by the surface charge density  $\sigma$ .

c) How does the field calculated in part *b* compare with the strength of the electric field a distance R from a point charge Q?

## **Exercise 4**

Repeat Exercise 1 assuming that the conducting sphere has a net charge of -Q. Does the charge on the conducting sphere have any effect on the fields inside the sphere? Why is there no field outside the sphere?

# **VAN DE GRAAFF GENERATOR**

The Van de Graaff generator is a conceptually straightforward device designed to produce high voltages. A sketch of the apparatus is shown in Figure (12), where we have a hollow metal sphere with a hole in the bottom, and a conveyer belt whose purpose is to bring charge up into the sphere. The belt is driven by a motor at the bottom.

The first step is to get electric charge onto the belt. This is done electrostatically by having an appropriate material rub against the belt. For example, if you rub a rubber rod with cat fur, you leave a negative charge on the rubber rod. If you rub a glass rod with silk, a positive charge will be left on the glass rod. I do not know what sign of charge is left on a comb when you run it through your hair on a dry day, but enough charge can be left on the comb to pick up small pieces of paper. We will leave the theory of creating electrostatic charges to other texts. For our discussion, it is sufficient to visualize that some kind of rubbing of the belt at the bottom near the motor deposits charge on the belt. (As an example of charging by rubbing, run a comb through your hair several times. The comb becomes electrically charged and will pick up small pieces of paper.)



## Figure 12

The Van de Graaff generator. Electric charge is carried up the belt and dumped inside the hollow metal sphere. Since there are no electric fields inside the sphere, the electric charge freely flows off the belt to the sphere, where it then spreads evenly to the <u>outside</u> surface of the sphere. Acting like a conveyor belt, the motorized belt carries the charge up and into the inside of the hollow metal sphere. If there is already charge on the sphere, then, as we have seen in Example (1), there will be an electric field outside the sphere as shown in Figure (13). (For this example we are assuming that the belt is carrying positive charge.) But inside the sphere there will be no field. (The hole in the bottom of the sphere lets a small amount of electric field leak inside, but not enough to worry about.)

As the charge is being carried up by the belt, the electric field outside the sphere pushes back on the charge, and the belt has to do work to get the charge up to the sphere. The more charge that has built up on the sphere, the stronger the electric field  $\vec{E}$ , and the more work the belt has to do. In a typical Van de Graaff generator used in lecture demonstration, you can hear the motor working harder when a large charge has built up on the sphere.



## Figure 13

It takes work to carry the charge up to the sphere against the electric field that is pushing down on the charge. But once inside the sphere where there is almost no field, the charge freely moves off the belt, onto the wire, charging up the sphere. The more charge on the sphere, the stronger the electric field  $\mathbf{E}$ outside the sphere, and the more work required to bring new charge up into the sphere. (In the demonstration model, you can hear the motor slow down as the sphere becomes charged up.) When the charge gets to the sphere how do we get it off the belt onto the sphere? When the sphere already has a lot of positive charge on it, why would the positive charge on the belt want to flow over to the sphere? Shouldn't the positive charge on the belt be repelled by the positive sphere?

Here is where our knowledge of electric fields comes in. As illustrated in Figure (13), there may be very strong electric fields outside the sphere, but inside there are none. Once the conveyor belt gets the charge inside the sphere, the charge is completely free to run off to the sphere. All we need is a small wire that is attached to the inside of the sphere that rubs against the belt. In fact, the neighboring + charge on the belt helps push the charge off the belt onto the wire.

Once the charge is on the wire and flows to the inside of the sphere, it must immediately flow to the outside of the sphere where it helps produce a stronger field  $\vec{E}$ shown in Figure (13).



#### Figure 14

We can discharge the Van de Graaff generator by bringing up a grounded sphere as shown. Since about 100,000 volts are required to make a spark one inch long, we can use the maximum length of sparks to estimate the voltage produced by the Van de Graaff generator.

## **Electric Discharge**

When a large amount of charge has accumulated on the metal sphere of the Van de Graaff generator, we can produce some very strong fields and high voltages. We can estimate the voltage by bringing a grounded sphere up to the Van de Graaff generator as shown in Figure (14). A voltage of about 100,000 volts is required to make a spark jump about an inch through air. Thus if we get a spark about 2 inches long between the Van de Graaff generator and the grounded sphere, we have brought enough charge onto the generator sphere to create a voltage of about 200,000 volts. (The length of the sparks acts as a crude voltmeter!)

As an exercise, let us estimate how many coulombs of charge must be on the Van de Graaff generator sphere to bring it up to a voltage of 200,000 volts.

Outside the Van de Graaff generator sphere, the electric field is roughly equal to the electric field of a point charge. Thus the voltage or electric potential of the sphere should be given by Equation (25-4) as

$$V = \frac{Q}{4\pi\varepsilon_0 r}$$
(25-4)

where r is the radius of the Van de Graaff generator sphere. (*Remember that r is not squared in the formula for potential energy or voltage.*)

Let us assume that r = 10 cm or .1 m, and that the voltage V is up to 200,000 volts. Then Equation (25-4) gives

$$Q = 4\pi\epsilon_0 rV$$
  
=  $4\pi \times 9 \times 10^{-12} \times .1 \times 200,000$   
$$Q \approx 2 \times 10^{-6} \text{ coulombs}$$

A couple millionth's of a coulomb of charge is enough to create 200,000 volt sparks. As we said earlier, a whole coulomb is a huge amount of charge!

## Grounding

The grounded sphere in Figure (14) that we used to produce the sparks, provides a good example of the way we use conductors and wires.

Beneath the Van de Graaff generator apparatus we have placed a large sheet of aluminum called a grounding plane that is attached to the metal pipes and the electrical ground in the room. (Whenever we have neglected to use this grounding plane during a demonstration we have regretted it.) We have attached a copper wire from the grounding plane to the "grounded" sphere as shown.

Thus in Figure (14), the grounding plane, the room's metal pipes and electrical ground wires, and the grounded sphere are all attached to each other via a conductor. Now there can be no electric field inside a conductor, therefore all these objects are at the same electric potential or voltage. (If you have a voltage difference between two points, there must be an electric field between these two points to produce the voltage difference.) It is common practice in working with electricity to define the voltage of the water pipes (or a metal rod stuck deeply into the earth) as zero volts or "ground". (The ground wires in most home wiring are attached to the water pipes.) Any object that is connected by a wire to the water pipes or electrical ground wire is said to be grounded. The use of the earth as the definition of the zero of electric voltage is much like using the floor of a room as the definition of the zero of the gravitational potential energy of an object.

In Figure (14), when the grounded sphere is brought up to the Van de Graaff generator and we get a 2 inch long spark, the spark tells us that the Van de Graaff sphere had been raised to a potential of at least 200,000 volts above ground.

Van de Graaff generators are found primarily in two applications. One is in science museums and lecture demonstration to impress visitors and students. The other is in physics research. Compared to modern accelerators, the 200,000 volts or up to 100 million volts that Van de Graaff generators produce is small. But the voltages are very stable and can be precisely controlled. As a result the Van de Graaff's make excellent tools for studying the fine details of the structure of atomic nuclei.

# THE ELECTRON GUN

In Figure (15) we have a rough sketch of a television tube with an electron gun at one end to create a beam of electrons, deflection plates to move the electron beam, and a phosphor screen at the other end to produce a bright spot where the electrons strike the end of the tube.

Figure (16) illustrates how a picture is drawn on a television screen. The electron beam is swept horizontally across the face of the tube, then the beam is moved down one line and swept horizontally again. An American television picture has about 500 horizontal lines in one picture.

As the beam is swept across, the brightness of the spot can be adjusted by changing the intensity of the electron beam. In Figure (16), line 3, the beam starts out bright, is dimmed when it gets to the left side of the letter A, shut off completely when it gets to the black line, then turned on to full brightness to complete the line. In a standard television set, one sweep across the tube takes about 60 microseconds. To draw the fine details you see on a good television set requires that the intensity of the beam can be turned up and down in little more than a tenth of a microsecond.



#### Figure 15

Cathode ray tubes, like the one shown above, are commonly used in television sets, oscilloscopes, and computer monitors. The electron beam (otherwise known as a "cathode ray") is created in the electron gun, is aimed by the deflection plates, and produces a bright spot where it strikes the phosphor screen.

The heart of this system is the electron gun which creates the electron beam. The actual electron gun in a television tube is a complex looking device with indirect heaters and focusing rings all mounted on the basic gun. What we will describe instead is a studentbuilt gun which does not produce the fine beam of a commercial gun, but which is easy to build and easy to understand.



#### Figure 16

The letter A on a TV screen. To construct an image the electron beam is swept horizontally, and turned up where the picture should be bright and turned down when dark. The entire image consists of a series of these horizontal lines, evenly spaced, one below the other.



#### Figure 17

Source of the electrons. The tungsten filament is heated by an electric current. When it becomes red-hot, electrons boil out through the surface. The white coating on the filament makes it easier for the electrons to escape.

## The Filament

As shown in Figure (17), the source of the electrons in an electron gun is the **filament**, a piece of wire that has been heated red-hot by the passage of an electric current. At these temperatures, some of the electrons in the filament gain enough thermal kinetic energy to evaporate out through the surface of the wire. The white coating you may see on a filament reduces the amount of energy an electron needs to escape out through the metal surface, and therefore helps produce a more intense beam of electrons. At standard temperature and pressure, air molecules are about 10- molecular diameters apart as indicated in Figure (18). Therefore if the filament is in air, an electron that has evaporated from the filament can travel, at most, a few hundred molecular diameters before striking an air molecule. This is why the red-hot burner on an electric stove does not emit a beam of electrons. The only way we can get electrons to travel far from the filament is to place the filament in a vacuum as we did in Figure (17). The better the vacuum, the farther the electrons can travel.



#### Figure 18

Whenever we heat a metal to a high enough temperature, electrons boil out of the surface. But if there is air at standard pressure around, the electrons do not get very far before striking an air molecule.

## **Accelerating Field**

Once the electrons are out of the filament we use an electric field to accelerate them. This is done by placing a metal cap with a hole in the end over the end of the filament as shown in Figure (19). The filament and cap are attached to a battery as shown in Figure (20) so that the cap is positively charged relative to the filament.

Intuitively the gun works as follows. The electrons are repelled by the negatively charged filament and are attracted to the positively charged cap. Most of the electrons rush over, strike, and are absorbed by the cap as shown in Figure (21). But an electron headed for the hole in the cap discovers too late that it has missed the cap and goes on out to form the electron beam.

A picture of the resulting electron beam is seen in Figure (22). The beam is visible because some air remains inside the tube, and the air molecules glow when they are struck by an electron.



Figure 19

To create a beam of electrons, we start by placing a metal cap with a hole in it, over the filament.



Figure 20

We then attach a battery to the metal cap so that the cap has a positive voltage relative to the filament.

# **A Field Plot**

A field plot of the electric field lines inside the electron gun cap gives a more precise picture of what is happening. Figure (23) is a computer plot of the field lines for a cylindrical filament inside a metal cap. We chose a cylindrical filament rather than a bent wire filament because it has the cylindrical symmetry of the cap and is therefore much easier to calculate and draw. But the fields for a wire filament are not too different.

First notice that the field lines are perpendicular to both metal surfaces. This agrees with our earlier discussion that an electric field at the surface of a conductor cannot have a parallel component for that would move the charge in the conductor. The second thing to note is that due to the unfortunate fact that the charge on the electron is negative, the electric field points oppositely to the direction of the force on the electrons. The force is in the direction of  $-\vec{E}$ .



## Figure 21

Electrons flow from the negative filament to the positive cap. The beam of electrons is formed by the electrons that miss the cap and go out through the hole.



**Figure 22** *Resulting electron beam.* 

The electrons, however, do not move along the  $-\vec{E}$  field lines. If they boil out of the filament with a negligible speed they will start moving in the direction  $-\vec{E}$ . But as the electrons gain momentum, the force  $-\vec{eE}$  has less and less effect. (Remember, for example, that for a satellite in a circular orbit, the force on the satellite is down toward the center of the earth. But the satellite moves around the earth in an orbit of constant radius.) In Figure (23), the dotted lines show a computer plot of the trajectories of the electrons at several points. The most important trajectories for our purposes are those that pass through the hole in the cap and go out and form the electron beam.

#### **Exercise 5**

Describe two other examples where an object does not move in the direction of the net force acting on it.





Plot of the electric field in the region between the filament and the cap. Here we assume that we have a cylindrical filament heated by a wire inside.

## **Equipotential Plot**

Once we know the field lines, we can plot the equipotential lines as shown in Figure (24). The lines are labeled assuming that the filament is grounded (0 volts) and that the cap is at 100 volts. The shape of the equipotentials, shown by dashed lines, does not change when we use different accelerating voltages, only the numerical value of the equipotentials changes.

The reason that the equipotential lines are of such interest in Figure (24) is that they can also be viewed as a map of the electron's kinetic energy.

Remember that the voltage V is the potential energy of a unit positive test charge. A charge q has a potential energy qV, and an electron, with a charge -e, has an electric potential energy -eV.

In our electron gun, the electrons evaporate from the filament with very little kinetic energy, call it zero. By the time the electrons get to the 10-volt equipotential, their electric potential energy has dropped to  $(-e \times 10)$  joules, and by conservation of energy, their kinetic energy has gone up to  $(+e \times 10)$  joules. At the 50 volt equipotential the electron's kinetic energy has risen to  $(e \times 50)$  joules, and when the electrons reach the 100 volt cap, their energy is up to  $(e \times 100)$  joules. Thus the equipotential lines in Figure (24) provide a map of the kinetic energy of the electrons.



#### Figure 24

Equipotential plot. We see that by the time the electrons have reached the hole in the cap, they have crossed the same equipotential lines and therefore have gained as much kinetic energy as the electrons that strike the cap. (From a student project by Daniel Leslie and Elad Levy.)
# ELECTRON VOLT AS A UNIT OF ENERGY

What is perhaps most remarkable about the electron gun is that every electron that leaves the filament and strikes the cap gains precisely the same kinetic energy. If we use a battery that produces 100 volt accelerating voltage, then every electron gains precisely ( $e \times 100$ ) joules of kinetic energy. This is also true of the electrons that miss the cap and go out and form the electron beam.

The amount of energy gained by an electron that falls through a 1 volt potential is  $(e \times 1 \text{ volt}) = 1.6 \times 10^{-19}$  joules. This amount of energy is called an *electron volt* and designated by the symbol eV.

$$1eV = \begin{cases} energy gained by an electron \\ falling through a 1 volt potential \\ = (e coulombs) \times (1 volt) \qquad (3) \\ = 1.6 \times 10^{-19} \text{ joules} \end{cases}$$

The dimensions in Equation (3) make a bit more sense when we realize that the volt has the dimensions of joule/coulomb, so that

$$1eV = (e \text{ coulombs}) \times \left(1 \frac{\text{joule}}{\text{coulomb}}\right)$$
$$= (e) \text{ joules}$$
(3a)

The electron volt is an extremely convenient unit for describing the energy of electrons produced by an electron gun. If we use a 100 volt battery to accelerate the electrons, we get 100 eV electrons. Two hundred volt batteries produce 200 eV electrons, etc.

To solve problems like calculating the speed of a 100 eV electron, you need to convert from eV to joules. The conversion factor is

$$\left| 1.6 \times 10^{-19} \frac{\text{joules}}{\text{eV}} \right| \qquad \begin{array}{c} conversion \\ factor \end{array}$$
(4)

For example, if we have a 100 eV electron, its kinetic energy  $1/2 \text{ mv}^2$  is given by

KE = 
$$1/2 \text{ mv}^2$$
  
=  $100 \text{ eV} \times 1.6 \times 10^{-19} \frac{\text{joules}}{\text{eV}}$  (5)

Using the value  $m = 9.11 \times 10^{-31}$  kg for the electron mass in Equation (5) gives

$$v = \sqrt{\frac{2 \times 100 \times 1.6 \times 10^{-19}}{9.11 \times 10^{-31}}}$$

$$= 6 \times 10^{6} \frac{\text{meters}}{\text{sec}}$$
(6)

which is 2% the speed of light.

In studies involving atomic particles such as electrons and protons, the electron volt is both a convenient and very commonly used unit. If the electron volt is too small, we can measure the particle energy in MeV (millions of electron volts) or GeV (billions of electron volts or *Gigavolts*).

$$1 \text{ MeV} \equiv 10^{6} \text{ eV}$$
$$1 \text{ GeV} \equiv 10^{9} \text{ eV}$$
(6)

For example, if you work the following exercises, you will see that the rest energies  $m_0c^2$  of an electron and a proton have the values

The reason that it is worth remembering that an electron's rest energy is about .5 MeV and a proton's about 1 GeV, is that when a particle's kinetic energy gets up toward its rest energy, the particle's speed becomes a significant fraction of the speed of light and nonrelativistic formulas like  $1/2 \text{ mv}^2$  for kinetic energy no longer apply.

# Example

Calculate the rest energy of an electron in eV.

## Solution:

$$E = \frac{m_0 c^2 \text{ joules}}{1.6 \times 10^{-19} \frac{\text{joules}}{\text{eV}}}$$
$$= \frac{9.11 \times 10^{-31} \times (3 \times 10^8)^2}{1.6 \times 10^{-19}}$$
$$= .51 \times 10^6 \text{ eV}$$

## Exercise 6

Calculate the rest energy of a proton in eV and GeV.

# Exercise 7

What accelerating voltage must be used in an electron gun to produce electrons whose kinetic energy equals their rest energy?



## Figure 24a

Another field plot by Leslie and Levy, showing the electric field and equipotential lines in a gun with a shorter cap.

# **About Computer Plots**

One final note in our discussion of the electron gun. You might feel that by using the computer plots in Figures (23) and (24) we have cheated a bit. We haven't done the work ourselves, we let somebody (or something) else do the calculations for us and we are just using their answers. Yes and no!

First of all, with a little bit of practice you can learn to draw sketches that are quite close to the computer plots. Use a trick like noting that field lines must be perpendicular to the surface of a conductor where they touch the conductor. If two conductors have equal and opposite charge—if they were charged by a battery—all the field lines that start on the positive conductor will stop on the negative one. Use any symmetry you can find to help sketch the field lines and then sketch the equipotential lines perpendicular to the field lines. Some places it is easier to visualize the equipotential lines, e.g., near the surface of a conductor, and then draw in the perpendicular field lines.

The other point is that, for a number of practical problems the geometry of the conductors is complicated enough that only by using a computer can we accurately plot the field lines and equipotentials. But once a computer plot is drawn, we do not have to worry about how it was calculated. Like a hiker in a new territory, we can use the computer plot as our contour map to tell us the shape and important features of the terrain. For example in our field plots of the electron gun, we see that there is virtually no field out in front of the hole where the electrons emerge, therefore from the time the electrons leave the hole they coast freely at constant speed and energy down the tube.

# THE PARALLEL PLATE CAPACITOR

Our final example in this chapter of fields and conductors is the parallel plate capacitor. Here we will work with a much simpler field structure than for the electron gun, and will therefore be able to calculate field strengths and voltages. The parallel plate capacitor serves as the prototype example of a capacitor, a device used throughout physics and electrical engineering for storing electric fields and electric energy.

Suppose we take two circular metal plates of area A, separate them by a distance d, and attach a battery as shown in Figure (25). This setup is called a parallel plate capacitor, and the field lines and equipotential for this setup are shown in the computer plot of Figure (26).

Except at the edges of the plates, the field lines go straight down from the positive to the negative plate, and the equipotentials are equally spaced horizontal lines parallel to the plates. If the plate separation d is small compared to the diameter D of the plates, then we can neglect the fringing of the field at the edge of the plates. The result is what we will call an ideal parallel plate capacitor whose field structure is shown in Figure (27). The advantage of working with this ideal capacitor is that we can easily derive the relationship between the charging voltage V, and the charge Q.



**Figure 25** *The parallel plate capacitor. The capacitor is charged up by connecting a battery across the plates as shown.* 



**Figure 26** *The electric field between and around the edge of the capacitor plates.* 

Let us take a close look at what we have in Figure (27). The electric field lines  $\vec{E}$  leave the positively charged top plate and go straight down to the negatively charged bottom plate. Since all the lines starting at the top plate stop at the bottom one, there must be an equal and opposite charge +Q and -Q on the two plates. There is no net charge on the capacitor, only a separation of charge. And because the field lines go straight down, nowhere do they get closer together or farther apart, the field must have a uniform strength E between the plates.

We can use Gauss' law to quickly calculate the field strength E. The top plate has a charge Q, therefore the total flux out of the top plate must be  $\Phi = Q/\epsilon_0$ . But we also have a field of strength E flowing out of a plate of area A. Thus flux of E flowing between the plates is  $\Phi = EA$ . Equating these two formulas for flux gives

$$\Phi = EA = \frac{Q}{\varepsilon_0}$$

$$E = \frac{Q}{\varepsilon_0 A}$$
(8)

We can relate the voltage V and the field strength E by remembering that *E* is the force on a unit test charge and V is the potential energy of a unit test charge. If I lift a unit positive test charge from the bottom plate a distance d up to the top one, I have to exert an upward force of strength E for a distance d and therefore do an amount of work  $E \times d$ . This work is stored as the electric potential energy of the unit test charge, and is therefore the voltage V:







In our idealized parallel plate capacitor the field lines go straight from the positive to the negative plate, and the field is uniform between the plates. It may seem surprising, but V is also the voltage of the battery (see Figure 25) used to charge up the capacitor.

There is also a simple relationship between the charge Q on the capacitor plates and the voltage difference V between them. Substituting the value of E from Equation (8) into Equation (9) gives

$$V = \left(\frac{d}{\varepsilon_0 A}\right) Q \tag{10}$$

Equation (10) makes an interesting prediction. If we have a fixed charge Q on the capacitor (say we charged up the capacitor and removed the battery), then if we increase the separation d between the plates, the voltage V will increase.

One problem with trying to measure this increase in voltage is that if we attach a common voltmeter between the plates to measure V, the capacitor will quickly discharge through the voltmeter. In order to see this effect we must use a special voltmeter called an *electrometer* that will not allow the capacitor to discharge. The classic electrometer, used in the 1800's, is the gold leaf electrometer shown in Figures (28) and (29). When the top plate of the electrometer is charged, some of the charge flows to the gold leaves, forcing the leaves apart. The greater the voltage, the greater the charge and the greater the force separating the leaves. Thus the separation of the leaves is a rough measure of the voltage.

In Figure (28), we see a gold leaf electrometer attached to two metal capacitor plates. When the plates are charged, the gold leaves separate, indicating that there is a voltage difference between the plates.

In Figures (29a,b), we are looking through the electrometer at the edge of the capacitor plates. In going from (29a) to (29b), we moved the plates apart without changing the charge on the plates. We see that when the plates are farther apart, the gold leaves are more separated, indicating a greater voltage as predicted by equation (10).



**Figure 28** Gold leaf electrometer attached to a parallel plate capacitor.



**Figure 29a** Looking through the electrometer at the edge of the charged capacitor plates.



#### Figure 29b

Without changing the charge, the plates are moved further apart. The increased separation of the gold leaves shows that the voltage difference between the capacitor plates has increased.

#### **Exercise 8**

Two circular metal plates of radius 10 cm are separated by microscope slide covers of thickness d = .12 mm. A voltage difference of 5 volts is set up between the plates using a battery as shown in Figure (25). What is the charge Q on the plates?

#### **Deflection Plates**

A fitting conclusion to this chapter is to see how the fields in parallel plate capacitor can be used to deflect the beam of electrons produced by an electron gun. In Figure (30) the beam of electrons from an electron gun is aimed between the plates of a parallel plate capacitor. The upward directed electric field  $\vec{E}$  produces a downward directed force -e $\vec{E}$  on the electrons, so that when the electrons emerge from the plates, they have been deflected downward by an angle  $\theta$  as shown. We wish to calculate this angle  $\theta$  which depends on the strength of the deflection voltage  $V_p$ , the length D of the plates, and on the speed v of the electrons.

While the electrons are between the plates, their acceleration is given by

$$\vec{A} = \frac{\vec{F}}{m_e} = \frac{-e\vec{E}}{m_e}$$

where  $m_e$  is the electron mass. This acceleration is constant and directed downward, just as in our old projectile motion studies. Using Equation (9) E = V/dfor the magnitude of  $\vec{E}$ , we find that the downward acceleration  $\vec{A}$  of the electrons has a magnitude



#### Figure 30

To deflect the beam of electrons, we place what is essentially a parallel plate capacitor in the path of the beam as shown. The electrons are deflected by the electric field between the capacitor plates.

$$\left| \vec{A} \right| = \frac{eE}{m_e} = \frac{eV_p}{m_e d}$$

where  $V_{p_{1}}$  is the voltage and d the separation of the deflection plates.

If a particle is subjected to a downward acceleration for a time T, and initially has no downward velocity, its final downward velocity  $v_{fy}$  is from the constant acceleration formulas as

$$v_{\rm fy} = A_{\rm y}T = \frac{eV_{\rm p}}{m_{\rm e}d} T$$
(11)

If the electrons emerge from the electron gun at a speed v, then the time T it takes them to pass between the plates is

$$T = \frac{D}{V}$$
(12)

The tangent of the deflection angle  $\theta$  is given by the ratio  $v_{fy}/v$  which we can get from Equations (11) and (12):

$$v_{fy} = \frac{eV_p}{m_e d} \times \frac{D}{v}$$
  $v_{fy}$ 

$$\tan \theta = \frac{v_{fy}}{v} = \frac{eV_p D}{m_e dv^2}$$
(13)

The final step is to note that the speed v of the electrons is determined by the electron gun accelerating voltage  $V_{acc}$  by the relationship

$$\frac{1}{2}m_{\rm e}v^2 = eV_{\rm acc} \quad or \quad v^2 = \frac{2eV_{\rm acc}}{m_{\rm e}} \tag{14}$$

Equations (13) and (14) finally give

$$\tan\theta = \frac{eV_pD}{m_ed(2eV_{acc}/m_e)} = \frac{1}{2}\frac{D}{d}\frac{V_p}{V_{acc}} \quad (15)$$

which is a fairly simple result considering the steps we went through to get it. It is reassuring that  $\tan\theta$  comes out as a dimensionless ratio, which it must.

#### **Exercise 9**

In an electron gun, deflection plates 5 cm long are separated by a distance d = 1.2 cm. The electron beam is produced by a 75 volt accelerating voltage. What deflection voltage V<sub>p</sub> is required to bend the beam 10 degrees?

#### Exercise 10

In what is called the Millikan oil drop experiment, shown in Figure (31), a vapor of oil is sprayed between two capacitor plates and the oil drops are electrically charged by radioactive particles.

Consider a particular oil drop of mass m that has lost one electron and therefore has an electric charge q = + e. (The mass m of the drop was determined by measuring its terminal velocity in free fall in the air. We will not worry about that part of the experiment, and simply assume that the drop's mass m is known.) To measure the charge q on the oil drop, and thus determine the electron charge e, an upward electric field  $\vec{E}$  is applied to the oil drop. The strength of the field E is adjusted until the upward electric force just balances the downward gravitational force. When the forces are balanced, the drop, seen through a microscope, will be observed to come to rest due to air resistance.

The electric field  $\vec{E}$  that supports the oil drop is produced by a parallel plate capacitor and power supply that can be adjusted to the desired voltage V. The separation between the plates is d.

a) Reproduce the sketch of Figure (31), Then put a + sign beside the positive battery terminal and a - sign beside the negative one.

b) Find the formula for the voltage V required to precisely support the oil drop against the gravitational force. Express your answer in terms of the geometry of the capacitor (plate separation d, area A, etc.) the drop's mass m, the acceleration due to gravity g, and the electron charge e.



**Figure 31** *Millikan oil drop apparatus* 

# Chapter 27 Basic Electric Circuits

In the modern age (post 1870) we have been surrounded by electric circuits. House wiring is our most familiar example, but we have become increasingly familiar with electric circuits in radio and television sets, and even the digital watch you may be wearing. In this chapter we will discuss the basic electric circuits in order to introduce the concepts of electric current, resistance, and voltage drops around the circuit. We will restrict ourselves to devices like batteries, resistors, light bulbs, and capacitors. The main purpose is to develop the background needed to work with electric circuits and electronic measuring equipment in the laboratory.

# **ELECTRIC CURRENT**

An electric current in a wire is conceptually somewhat like the current of water in a river. We can define the current in a river as the amount of water per second flowing under a bridge. The amount of water could be defined as the number of water molecules, but a more convenient unit would be gallons, liters, or cubic meters.

An electric current in a wire is usually associated with the flow of electrons and is measured as the amount of charge per second flowing past some point or through some cross-sectional area of the wire, as illustrated in Figure (1). We could measure the amount of charge by counting the number of electrons crossing the area, but it is more convenient to use our standard unit of charge, the coulomb, and define an electric current as the number of coulombs per second passing the crosssectional area. The unit of current defined this way is called an *ampere*.

$$1 \text{ ampere} = \begin{cases} 1 \text{ coulomb per} \\ \text{second passing} \\ \text{a cross-sectional} \\ \text{area of wire} \end{cases}$$
(1)

From your experience with household wiring you should already be familiar with the ampere (amp) as a unit of current. A typical light bulb draws between 1/2 and 1 ampere of current, and so does the typical motor in an electric appliance (drill, eggbeater, etc.). A microwave oven and a toaster may draw up to 6 amps, and hair dryers and electric heaters up to 12 amps. Household wiring is limited in its capability of carrying electric current. If you try to carry too much current in a wire, the wire gets hot and poses a fire hazard.



**Figure 1** An electric current is defined as the amount of <u>charge</u> <u>per second</u> flowing past a cross-sectional area.

Household wiring is protected by fuses or circuit breakers that shut off the current if it exceeds 15 or 20 amps. (You can see why you do not want to run a hair dryer and an electric heater on the same circuit.)

There is a common misconception that the electrons in a wire travel very fast when a current is flowing in the wire. After all when you turn on a wall light switch the light on the other side of the room appears to turn on instantly. How did the electrons get there so fast?

The answer can be seen by an analogy to a garden hose. When you first attach an empty hose to a spigot and turn on the water, it takes a while before the hose fills up with water and water comes out of the other end. But when the hose is already full and you turn on the spigot, water almost instantly comes out of the other end. Not the water that just went in, but the water that was already in the hose.

A copper wire is analogous to the hose that is already full of water; the electrons are already there. When you turn on the light switch, the light comes on almost instantly because all the "electric fluid" in the wire starts moving almost at once.

To help build an intuition, let us estimate how fast the electrons must move in a copper wire with a 1 millimeter cross-sectional area carrying an electric current of one ampere. This is not an unreasonable situation for household wiring.

A copper atom has a nucleus containing 29 protons surrounded by a cloud of 29 electrons. Of the 29 electrons, 27 are tightly bound to the nucleus and 2 are in an outer shell, loosely bound. (All metal atoms have one, two, and sometimes 3 loosely bound outer electrons.) When copper atoms are collected together to form a copper crystal, the 27 tightly bound electrons remain with their respective nuclei, but the two loosely bound electrons are free to wander throughout the crystal. In a metal crystal or wire, it is the loosely bound electrons (called *conduction electrons*) that form the electric fluid that makes the wire a conductor. Copper has an atomic weight of 63.5, thus there are 63.5 grams of copper in a mole. And the density of copper is 9 gm/cm<sup>3</sup>, thus a mole of copper has a volume

volume of one  
mole of copper 
$$\left. \right\} = \frac{63.5 \text{gm} / \text{mole}}{9 \text{gm} / \text{cm}^3} = 7 \frac{\text{cm}^3}{\text{mole}}$$

Since a mole of a substance contains an Avogadro's number  $6 \times 10^{23}$  of particles of that substance, and since there are 2 conduction electrons per copper atom, 7 cm<sup>3</sup> of copper contain  $12 \times 10^{23}$  conduction electrons. Dividing by 7, we see that there are  $1.7 \times 10^{23}$  conduction electrons in every cubic centimeter of copper and  $1.7 \times 10^{20}$  in a cubic millimeter. Converting this to coulombs, we get



= 
$$27 \text{ coulombs/mm}^3$$

In our 1 millimeter cross-sectional area wire, if the electrons flowed at a speed of 1 millimeter per second, 27 coulombs of charge would flow past any point in the wire per second, and we would have a current of 27 amperes. To have a current of 1 ampere, the electrons would have to move only 1/27 as fast, *or 1/27 of a millimeter per second*! This slow speed results from the huge density of conduction electrons.

#### **Positive and Negative Currents**

If you are using a hose to fill a bucket with water, there is not much question about which way the current of water is flowing—from the hose to the bucket. But with electric current, because there are two kinds of electric charge, the situation is not that simple. As shown in Figure (2), there are two ways to give an object a positive charge, add positive charge or remove negative charge. If a wire connected to the object is doing the charging, it may be difficult to tell whether there is a current of positive charge into the object or a current of negative charge out of the object. Both have essentially the same effect.

You may argue that at least for copper wires a current of positive charge doesn't make sense because the electric current is being carried by the negative conduction electrons. But a simple model of an electric current will clearly demonstrate that a positive current flowing one way is essentially equivalent to a negative current flowing the other way.



A current of positive charge into an object, or a current of negative charge out, leaves the object positively charged.

In Figure (3) we have tried to sketch a picture of a copper wire in which the conduction electrons are moving to the left producing a left directed negative current. The problem with Figure (3) is that it is hard to show the conduction electrons flowing through the lattice of stationary positive copper nuclei. The picture is difficult to draw, and Figure (3) is not particularly informative.

To more clearly show that the positive charge is at rest and that it is the negative charge that is moving, we have in Figure (4) constructed a model of a copper wire in which we have two separate rods, one moving and one at rest. The stationary rod has the positive copper nuclei and the moving rod has the negative conduction electrons. This model is not a very good representation of what is going on inside the copper wire, but it does remind us clearly that the positive charge is at rest, and that the current is being carried by the moving negative charge.

When you see this model, which we will use again in later discussions, think of the two rods as merged together. Picture the minus charge as flowing through the lattice of positive charge. Remember that the only reason that we drew them as separate rods was to clearly show which charge was carrying the electric current.

Using the results of the previous section, we can make our model of Figure (4) more specific by assuming it represents a copper wire with a 1 millimeter cross section carrying a current of one ampere. In that



## Figure 3

A copper wire at rest with the conduction electrons moving to the left. This gives us a left-directed negative current. example the average speed of the conduction electrons was 1/27 of a millimeter per second, which we will take as the speed v of the moving negative rod in Figure (4).

Figure (5a) is the same as Figure (4), except we have drawn a stick figure representing a person walking to the left at a speed v. The person and the negatively charged rod are both moving to the left at the same speed.

Figure (5b) is the same situation from the point of view of the stick figure person. From her point of view, the negative rod is at rest and it is the positive rod that is moving to the right. *Our left directed negative current in Figure (5a) is seen by the moving observer to be a right directed positive current (Figure 5b).* Whether we have a left directed negative current or a right directed positive current just depends upon the point of view of the observer.

But how fast was our moving observer walking? If Figure (5) is a model of a 1 mm<sup>2</sup> copper wire carrying a current of 1 ampere, *the speed v in Figure (5) is 1/27 of a millimeter per second*. This is about 2 millimeters per minute! Although faster than the continental drift, this motion should certainly have little effect on what we see. If the wire is leading to a toaster, the toast will come out the same whether or not we walk by at a speed of 2 mm per minute. For most purposes, we can take a left directed negative current and a right directed positive current as being equivalent. Relatively sophisticated experiments, such as those using the Hall effect (to be discussed later) are required to tell the difference.



Model of a copper wire carrying an electric current. We are representing the positive copper ions by a positively charged rod at rest, and the conduction electrons by a moving, negatively charged rod.

# **A Convention**

It was Ben Franklin who made the assignment of positive and negative charge. The charge left on a glass rod rubbed by silk was defined as positive, and that left on a rubber rod rubbed by cat fur as negative. This has often been considered a tragic mistake, for it leaves the electron, the common carrier of electric current, with a negative charge. It also leads to the unfortunate intuitive picture that an atom that has lost some electrons ends up with a positive charge.

Some physics textbooks written in the 1930s redefined the electron as being positive, but this was a disaster. We cannot undo over two centuries of convention that leads to the electron as being negative.

The worst problem with Franklin's convention comes when we try to handle the minus signs in problems involving the flow of electrons in a wire. But we have just seen that the flow of electrons in one direction is almost completely equivalent to the flow of positive charge in the other. If we do our calculations for positive currents, then we know that the electrons are simply moving in the opposite direction.

In order to maintain sanity and not get tangled up with minus signs, in this text we will, whenever possible, talk about the flow of positive currents, and talk about the force on positive test charges. If the problem we are working on involves electrons, we will work everything assuming positive charges and positive currents, and only at the end of the problem we will take into account the negative sign of the electron. With some practice, you will find this an easy convention to use.



a) observer walking along with the moving negatively charged rod



 b) from the observer's point of view the negative rod is at rest and the positive charge is moving to the right

## Figure 5 a, b

In (a) we have a left directed negative current, while in (b) we have a right directed positive current. The only difference is the perspective of the observer. (You can turn a negative current into an oppositely flowing positive one simply by moving your head.)

# **CURRENT AND VOLTAGE**

Students first studying electricity can have difficulty conceptually distinguishing between the concepts of current and voltage. This problem can be handled by referring back to our hydrodynamic analogy of Chapter 23.

In Chapter 23 we were discussing Bernoulli's equation which stated that the quantity  $(P + \rho gh + 1/2 \rho v^2)$  was constant along a stream line if we could neglect viscous effects in the fluid. Because of the special nature of this collection of terms, we gave them the name *hydrodynamic voltage*.

$$\frac{\text{hydrodynamic}}{\text{voltage}} = P + \rho gh + \frac{1}{2}\rho v^2 \quad (23-23)$$

(The second and third terms in the hydrodynamic voltage are the potential energy of a unit volume of fluid and the kinetic energy. The pressure term, while not a potential energy, is related to the work required to move fluid into a higher pressure region.)

Many features of hydrodynamic voltage should already be familiar. If you live in a house with good water pressure, when you turn on the faucet the water comes out rapidly. But if someone is running the washing machine in the basement or watering the garden, the water pressure may be low, and the water just dribbles out of the faucet. We will think of the high pressure water as *high voltage* water, and the low pressure water as *low voltage* water.

Let us look more carefully at high voltage water in a faucet. When the faucet is shut off, the water is at rest but the pressure is high, and the main contribution to the hydrodynamic voltage is the P term. When the faucet is on, the water that has just left the faucet has dropped back to atmospheric pressure but it is moving rapidly. Now it is the  $1/2 \rho v^2$  that contributes most to the hydrodynamic voltage. If the water originally comes

from a town water tank, when the water was at the top of the tank it was at atmospheric pressure and not moving, but was at a great height h. In the town water tank the hydrodynamic voltage comes mainly from the  $\rho$ gh term.

Let us focus our attention on the high pressure in a faucet that is shut off. In this case we have high voltage water but no current. We can get a big current if we turn the faucet on, but the voltage is there whether or not we have a current.

In household wiring, the electrical outlets may be thought of as faucets for the electrical fluid in the wires. The high voltage in these wires is like the high pressure in the water pipes. You can have a high voltage at the outlet without drawing any current, or you can connect an appliance and draw a current of this high pressure electrical fluid.

# Resistors

In an electric heater the electrical energy supplied by the power station is converted into heat energy by having electric current flow through a dissipative or resistive material. The actual process by which electrical energy is turned into heat energy is fairly complex but not unlike the conversion of mechanical energy to heat through friction. One can think of resistance as an internal friction encountered by the electric current.

In our discussion of Bernoulli's equation we saw that the hydrodynamic voltage  $P + \rho gh + 1/2 \rho v^2$  was constant along a stream line if there were no viscous effects. But we also saw in Figure (23-24) that when there were viscous effects this hydrodynamic voltage dropped as we went along a stream line.



**Figure 23-24** *Hydrodynamic voltage drop due to viscous effects.* 

In fluid flows, we get the most dissipation where the fluid is moving rapidly through a narrow constriction. This is seen in our venturi demonstration of Figure (23-18), reproduced here in Figure (6). Here we have a large tube with a constriction. The glass barometer tubes show us that the pressure remains relatively constant before the constriction, but does not return to its original value afterward. There is a net pressure drop of  $\rho$ gh, where h is the height drop indicated in the figure.

Consider the points in the fluid at the dots labeled (2) and (9), in the center of the stream below tubes 2 and 9. These points are at the same heights ( $h_2 = h_9$ ), and the fluid velocities are the same ( $v_2 = v_9$ ) because the flow tube has returned to its original size. Because of the pressure drop ( $P_9 < P_2$ ), the hydrodynamic voltage ( $P_9 + \rho g h_9 + 1/2 \rho v_9^2$ ) at point (9) is less than that at point (2) by an amount equal to  $P_2-P_9 = \rho gh$ . The barometer tubes 2 and 9 are acting as *hydrodynamic voltmeters* showing us where the voltage drop occurs.

Just as in fluid flows, dissipation in electric currents are associated with voltage drops, in this case electrical voltage drops. In general, the amount of the voltage drop depends on the amount of current, the geometry of the flow path, on the material through which the current is flowing, and on the temperature of the material. But in a special device called a *resistor*, the voltage drop  $\Delta V$ depends primarily on the current i through the resistor



Figure 6

The hydrodynamic voltage, as measured by the barometer tubes, drops by an amount  $\rho$ gh in going across the constriction from Point (2) to Point (9).

and is proportional to that current. When the voltage drop  $\Delta V$  is proportional to the current *i*, the resistor is said to obey Ohm's law. This can be written as the equation

$$\Delta V = iR \qquad Ohmis \, law \qquad (1)$$

The proportionality constant R is called the *resistance* R of the resistor. From Equation (1) you can see that R has the dimensions volt/amp. This unit is called an *ohm*, a name which is convenient in practice but which further complicates the problem of following dimensions in electrical calculations.

$$R = \frac{\Delta V}{i} \frac{volts}{amps} = \frac{\Delta V}{i} ohms$$

Resistors are the most common element in electronic circuits. They usually consist of a small cylinder with wire pigtails sticking out each end as shown in Figure (7). The material inside the cylinder which creates the voltage drop, which turns electrical energy into heat energy, is usually carbon.

The resistors you find in an electronics shop come in a huge selection of values, with resistances ranging from about 0.1 ohm up to around  $10^9$  ohms in a standard series of steps. The physical size of the resistor depends not on the value of the resistance but on the amount of electrical energy the resistor is capable of dissipating without burning up. The value of the resistance is usually indicated by colored stripes painted on the resistor, there being a standard color code so that you can read the value from the stripes.

(A light bulb is a good example of an electrical device that dissipates energy, in this case mostly in the form of heat and some light. The only problem with a light bulb is that as the filament gets hot, its resistance increases. If we wish to use Ohm's law, we have to add the qualification that the bulb's resistance R increases with temperature.)



#### Figure 7

The resistor, found in most electronic circuits. The purpose of the resistor is to cause an electric voltage drop analogous to the hydrodynamic voltage drop we saw in Figure 6 across the restriction in the flow tube.

## **A Simple Circuit**

To get some intuition for how resistors are used, consider the circuit shown in Figure (8) containing a battery and a resistor connected by wires. In drawing circuits, it is convention to use a line \_\_\_\_\_\_ for a wire, the symbol  $-\sqrt{\sqrt{\sqrt{\sqrt{\sqrt{\sqrt{-}}}}}}$  for a resistor, and  $-\frac{+}{|i^-|}$  for a battery. In the symbol for a battery, the short perpendicular line represents the negative terminal of the battery and the long side the positive terminal. When we have a current i flowing through the wire we draw an arrow indicating the direction of flow of positive charge  $-\frac{i}{2}$  and label the current with a letter such as i,  $i_1$ , etc.

In Figure (9), we have labeled the voltages  $V_1$ ,  $V_2$ ,  $V_3$  and  $V_4$  at four points around the circuit. By definition we will take the negative side of the battery as being zero volts, or what we call *ground* 

$$V_4 = 0$$
 volts by definition (2)

On the positive side of the battery, the voltage is up to the battery voltage  $V_b$  which is 1.5 volts for a common flashlight battery and up to 9 volts for many transistor radio batteries

$$V_1 = V_b$$
 the battery voltage (3)



#### Figure 8

About the simplest electrical circuit consists of a battery connected to a resistor. If the resistor were a light bulb, you would have a flashlight.

Point (2) at the upper end of the resistor, is connected to the positive terminal of the battery, Point (1), by a wire. In our circuit diagrams we always assume that our wires are good conductors, having no electric fields inside them and therefore no voltage drops along them. Thus

$$V_2 = V_1 (= V_b) \qquad \begin{array}{c} no \ voltage \ drop\\ along \ a \ wire \end{array}$$
(4)

The bottom of the resistor is connected to the negative terminal of the battery by a wire, therefore

$$V_3 = V_4 (=0) \qquad \begin{array}{c} no \ voltage \ drop\\ along \ a \ wire \end{array}$$
(5)

Equations (4) and (5) determine the voltage drop  $\Delta V$  that must be occurring at the resistor

$$\Delta V = V_2 - V_3 = V_b \tag{6}$$

And by Ohm's law, Equation (1), this voltage drop is related to the current i through the resistor by

$$\Delta V = iR = V_{\rm b} \qquad Ohms \, law \qquad (7)$$

Solving for the current i in the circuit gives

$$i = \frac{V_b}{R}$$
(8)

In future discussions of circuits we will not write out all the steps as we have in Equations (2) through (8), but the first time through a circuit we wanted to show all the details.



**Figure 9** *Voltages around the circuit.* 

Equation (8) is the one that really shows us how resistors are used in a circuit. We can see from Equation (8) that if we use a small resistor, we get a big current, and if we use a large resistor we get a small current. *In most applications resistors are used to control the flow of current*.

In modern electronics such as radios and computers, typical battery voltages are around 5 volts and typical currents a milliampere ( $10^{-3}$  amps). What size resistor R do we have to use in Equation (8) so that we get a one milliampere current from a 5 volt battery? The answer is

$$R = \frac{V_b}{i} = \frac{5 \text{ volts}}{10^{-3} \text{ amps}}$$
$$= 5000 \text{ ohms} \equiv 5000 \Omega$$
(9)

where we used the standard symbol  $\Omega$  for ohms. Many of the resistors in electronics circuits have values like this in the 1,000  $\Omega$  to 10,000  $\Omega$  range.

#### **The Short Circuit**

Equation (8) raises an interesting problem. What if R = 0? The equation predicts an infinite current! We could try to make R = 0 by attaching a wire rather than a resistor from Points (2) to (3) in Figure (9). What would happen is that a very large current would start to flow and either melt the wire, start a fire, drain the battery, or destroy the power supply. (A power supply is an electronic battery.) When this happens, you have created what is called a *short circuit*. The common lingo is that you have *shorted* out the battery or power supply and this is not a good thing to do.

#### Power

As one of the roles of a resistor is electrical power dissipation, let us determine the power that is being dissipated when a current is flowing through a resistor. Recall that power is the amount of energy transferred or dissipated per unit time. In the MKS system power has the dimensions of joules per second which is called a watt

$$Power = \frac{joules}{second} = watt$$
(10)

Now suppose we have a current flowing through a resistor R as shown in Figure (10). The voltage drop across the resistor is V, from a voltage of V volts at the top to 0 volts at the bottom as shown.

Because V is the electric potential energy of a unit charge (the coulomb), every coulomb of charge flowing through the resistor loses V joules of electric potential energy which is changed to heat.

If we have a current i, then i coulombs flow through the resistor every second. Thus the energy lost per second is the number of coulombs (i) times the energy lost per coulomb (V) or (iV):

Power = 
$$i \frac{coul}{sec} \times V \frac{joules}{coul}$$
  
=  $iV \frac{joules}{sec}$  =  $iV$  watts (11)

Ohm's law, Equation (1), can be used to express the power in terms of R and either i or V

Power = 
$$iV = i^2R = \frac{V^2}{R}$$
 (11a)

#### Figure 10

The voltage drops from V to 0 as the current i flows through the resistor. The power dissipated is the current i coulombs/second times the voltage drop V joules/coulomb, which is iV joules/second, or watts.

#### **Exercise 1**

These are some simple exercises to have you become familiar with the concepts of volts and amps.

a) Design a circuit consisting of a 9 volt battery and a resistor, where the current through the resistor is 25 milliamperes ( $25 \times 10^{-3}$  amps).

b) A flashlight consists of a 1.5 volt battery and a 1 watt light bulb. How much current flows through the bulb when the flashlight is on?

c) When you plug a 1000 watt heater into a 120 volt power line, how much current goes through the heater? What is the resistance R of the heater when the filament is hot?

d) In most households, each circuit has a voltage of 120 volts and is fused for 20 amps. (The circuit breaker opens up if the current exceeds 20 amps). What is the maximum power you can draw from one circuit in your house?

e) An electric dryer requires 3000 watts of power, yet it has to be plugged into wires that can handle only 20 amps. What is the least voltage you can have on the circuit?

f) In many parts of the world, the standard voltage is 240 volts. The wires to appliances are much thinner. Explain why.



**Figure 9 (redrawn)** *Voltages around the circuit.* 

## **KIRCHOFF'S LAW**

Imagine that you are going for an afternoon hike on a nearby mountain. You drive up to the base lodge, park your car, and start up the trail. The trail goes up over a ridge, down into a ravine, up to the peak of the mountain, down the other side and then around the mountain back to the base lodge. When you get back to your car, how much gravitational potential energy have you gained from the trip? The answer is clearly zero—you are right back where you started.

If you defined gh, which is the potential energy of a unit mass, as your gravitational voltage, then as you went up the ridge, there was a voltage rise as h increased. Going down into the ravine there was a voltage drop, or what we could call a negative voltage rise. The big voltage rise is up to the top of the mountain, and the big negative voltage rise is down the back side of the mountain. When you add up all the voltage rises for the complete trip, counting voltage drops as negative rises, the sum is zero.

Consider our Figure (9) redrawn here. If we start at Point (4) where the voltage is zero, and "walk" around the circuit in the direction of the positive current i, we first encounter a voltage rise up to  $V = V_h$  due to the battery, then a voltage drop back to zero at the resistor. When we get back to the starting point, the sum of the voltage rises is zero just as in our trip through the mountains. Even in more complicated circuits with many branches and different circuit elements, it is usually true that the sum of the voltage rises around any complete path, back to your starting point, is zero. It turns out that this is a powerful tool for analyzing electric circuits, and is known as Kirchoff's law. (Kirchoff's law can be violated, we can get a net voltage rise in a complete circuit, if changing magnetic fields are present. We will treat this phenomenon in a later chapter. For now we will discuss the usual situation where Kirchoff's law applies.)

# **Application of Kirchoff's Law**

There are some relatively standard, cookbook like procedures that make it easy to apply Kirchoff's law to the analysis of circuits. The steps in the recipe are as follows:

(1) Sketch the circuit and use arrows to show the direction of the positive current in each loop as we did in Figure (11). Do not be too concerned about getting the correct direction for the current i. If you have the



**Figure 11** Labeling the direction of the current.

arrow pointing the wrong way, then when you finish solving the problem, i will turn out to be negative.

(2) Label all the *voltage rises* in the circuit. Use arrows to indicate the direction of the voltage rise as we did in Figure (12). Note that if we go through the resistor in the direction of the current, we get a voltage drop. Therefore the arrow showing the voltage rise in a resistor must point back, opposite to the direction of the



current i in the resistor. (The analogy is to a rock strewn waterfall where the water loses hydrodynamic voltage as it flows down through the rocks. The direction of the voltage rise is back up the waterfall, in a direction opposite to that of the current.)

(3) The final step is to "*walk*" around the loop in the direction of i (or any direction you choose), and set the sum of the voltage rises you encounter equal to zero. If you encounter an arrow that points in the direction you are walking, it counts as a positive voltage rise (like  $V_b$  in Figure 12). If the arrow points against you (like  $V_R$ ), then it is a negative rise. Applying this rule to Figure (12) gives

Sum of the voltage rises  
going clockwise around  
the circuit of Figure 12 
$$= V_b + V_R$$
(12)  
$$= V_b + (-iR)$$

Equation (12) gives

$$i = \frac{V_b}{R}$$
(13)

= 0

which is the result we had back in Equation (8).

## **Series Resistors**

By now we have beaten to death our simple battery resistor circuit. Let us try something a little more challenging—let us put in two resistors as shown in Figure (13). In that figure we have drawn the circuit and labeled the direction of the current (Step 1), and drawn in the arrows representing the voltage rises (Step 2). Setting the sum of the voltage rises equal to zero (Step 3) gives

$$V_{b} + (-iR_{1}) + (-iR_{2}) = 0$$
(14)

$$i = \frac{V_b}{(R_1 + R_2)}$$
 (15)

The two resistors in Figure (13) are said to be connected in series. Comparing Equation (13) for a single resistor and Equation (15) for the series resistors, we see that if

$$R_1 + R_2 = R$$
 (series resistors) (16)

then we get the same current i in both cases (if we use the same battery). We say that if  $R_1 + R_2 = R$  then the series resistors are equivalent to the single resistor R.



**Figure 13** *Two resistors in series.* 

## **Parallel Resistors**

A bit more challenging is the circuit of Figure (14) where the resistors are wired in "parallel". In Step (1), we drew the circuit and labeled the currents. But here we have something new. When the current gets to the point labeled (A), it is like a fork in the stream and the current divides. We have labeled the two branch currents  $i_1$  and  $i_2$ , and have the obvious subsidiary condition (conservation of current, if you like).

$$i_1 + i_2 = i$$
 (17)

There is no problem with Step (2), the voltage rises are  $V_b$ ,  $i_1R_1$  and  $i_2R_2$  as shown. But we get something new when we try to write down Kirchoff's law for the sum of the voltage rises around a complete circuit. Now we have three different ways we can go around a complete circuit, as shown in Figures (15 a, b, c).

Applying Kirchoff's law to the path shown in Figure (15a) we get

$$V_b + (-i_1 R_1) = 0 (18)$$

For Figure (15b) we get

$$(-i_2 R_2) + (i_1 R_1) = 0 \tag{19}$$

and for Figure (15c) we get

$$V_b + (-i_2 R_2) = 0 (20)$$



**Figure 14** *Two resistors in parallel.* 

The main problem with using Kirchoff's laws for complex circuits is that we can get more equations than we need or want. For our current example, if you solve Equation (18) for  $V_b = i_1R_1$ , then put that result in Equation (20), you get  $i_1R_1 - i_2R_2 = 0$  which is Equation (19). In other words Equation (19) does not tell us anything that we did not already know from Equations (18) and (20). The mathematicians would say that Equations (18), (19), and (20) are not linearly independent.

Let us look at the situation from a slightly different point of view. To completely solve the circuit of Figure (15), we have to determine the currents i,  $i_1$  and  $i_2$ . We have three unknowns, but four equations, Equations (17), (18), (19) and (20). It is well known that you need as many equations as unknowns to solve a system of equations, and therefore we have one too many equations.





**Figure 15** Three possible loops for analyzing the parallel resistance circuit. They give more equations than needed.

We cannot arbitrarily throw out one of the equations for the remaining three must be linearly independent. For example, if we threw out Equation (17) and tried to solve Equations (18), (19) and (20) for  $i_{1}$ ,  $i_{2}$ , and i, we couldn't get an answer because Equation (19) contains no information not already in Equation (18) and (20). When you are working with a system of linear equations, the hardest problem is to decide which is a set of linear independent equations. Then you can use a standard set of procedures that mathematicians have for solving linear equations. These procedures involve determinants and matrices, which are easily handled on a computer, but are tedious to work by hand.

In our treatment of circuit theory we will limit our discussion to simple circuits where we can use grade school methods for solving the equations. Problems of linear independence, determinants and matrices will be left to other treatments of the topic.

To solve our parallel resistor circuit of Figure (14), we have from Equation (18)

$$i_1 = \frac{V_b}{R_1}$$

and from Equation (20)

$$i_2 = \frac{V_b}{R_2}$$

Substituting these values in Equation (17) gives

$$i = i_{1} + i_{2} = \frac{V_{b}}{R_{1}} + \frac{V_{b}}{R_{2}}$$

$$= V_{b} \left(\frac{1}{R_{1}} + \frac{1}{R_{2}}\right)$$
(21)

Comparing Equation (21) for parallel resistors, and Equation (13) for a single resistor

$$i = V_b \left(\frac{1}{R}\right) \tag{13}$$

We see that two parallel resistors  $R_1$  and  $R_2$  are equivalent to a single resistor R if they obey the relationship

$$\boxed{\frac{1}{R} = \frac{1}{R_1} + \frac{1}{R_2}} \qquad \begin{array}{c} equivalent \\ parallel \\ resistors \end{array} (22)$$

#### **Exercise 2**

You are given a device, sealed in a box, with electrical leads on each end. (Such a device is often referred to as a "black box", the word black referring to our lack of knowledge of the contents, rather than the actual color of the device.) You use an instrument called an **ohmmeter** to measure the electrical resistance between the two terminals and find that it's resistance R is 470 ohms ( $470\Omega$ ).



a) Sketch a circuit, containing the black box and one resistor, where the total resistance of the circuit is 500  $\Omega$  .

b) Sketch a circuit, containing the black box and one resistor, where the total resistance of the circuit is  $400\,\Omega$  .

#### Exercise 3 The Voltage Divider

We wish to measure the voltage  $V_b$  produced by a high voltage power supply, but our voltmeter has the limited range of +2 to -2 volts. To make the measurement we use the voltage divider circuit shown below, containing a big resistor  $R_1$  and a small resistor  $R_2$ . If, for example,  $R_2$  is 1000 times smaller than  $R_1$ , then the voltage across  $R_2$  is 1000 times smaller than that across  $R_1$ . By measuring the small voltage across the small resistor we can use this result to determine the big voltage  $V_b$ .

a) What current i flows through the circuit. Express your answer in terms of  $V_{b}$ .

b) Find the formula for  $V_b$  in terms of  $V_2$ , the voltage measured *across* the small resistor.

c) Find a formula for  $V_b$  in terms of  $V_2$ ,  $R_1$  and  $R_2$ , assuming  $R_1 > > R_2$ , so that you can replace  $(R_1 + R_2)$  by  $R_1$  in the equation for i.

d) Our voltmeter reads  $V_2 = .24$  volts. What was  $V_b$ ?



# **CAPACITANCE AND CAPACITORS**

In addition to the resistor, another common circuit element is the capacitor. A resistor dissipates energy, causes a voltage drop given by Ohm's law V = iR, and is often used to limit the amount of current flowing in a section of a circuit. A capacitor is a device for storing electrical charge and maintains a voltage proportional to the charge stored. We have already seen one explicit example of a capacitor, the parallel plate capacitor studied in the last chapter. Here we will abstract the general features of capacitors, and see how they are used as circuit elements.

## Hydrodynamic Analogy

Before focusing on the electrical capacitor, it is instructive to consider an accurate hydrodynamic analogy the cylindrical water tank shown in Figure (16). If the tank is filled to a height h, then all the water in the tank has a hydrodynamic voltage

$$V_{\rm h} = P + \rho g h + \frac{1}{2} \rho v^2 = \rho g h \qquad (23)$$

For water at the top of the tank, y = h, the voltage is all in the form of gravitational potential energy  $\rho gh$ . (We will ignore atmospheric pressure.) At the bottom of the tank where y = 0, the voltage is all in the pressure term  $P = \rho gh$ . The dynamic voltage term  $1/2 \rho v^2$  does not play a significant role.





Analogy between a cylindrical tank of water and an electrical capacitor. In the tank, all the water in the tank is at a hydrodynamic voltage  $V_h = \rho gh$ , and the quantity Q of water in the tank, given by  $Q = Ah = (A/\rho g)\rho gh = (A/\rho g)V_h$  is proportional to  $V_h$ .

Let us denote by the letter Q the quantity or volume of water stored in the tank. If we talk only about cylindrical tanks (of cross-sectional area A), then this volume is proportional to the height h and therefore the hydrodynamic voltage  $V_h$ 

Volume of water  
in cylindrical tank 
$$= Q = Ah = \left(\frac{A}{\rho g}\right)\rho gh$$
  
$$Q = \left(\frac{A}{\rho g}\right)V_h$$
 (24)

If we define the proportionality constant  $A/\rho g$  in Equation (24) as the *capacitance C* of the tank

$$C = \frac{A}{\rho g} \equiv \begin{cases} capacitance \text{ of} \\ a \text{ cylindrical tank} \\ \text{with a cross } - \\ \text{sectional area A} \end{cases}$$
(25)

then we get

$$Q = CV_{h}$$
(26)

as the relation between the hydrodynamic voltage and volume Q of water in the tank.

## Cylindrical Tank as a Constant Voltage Source

One of the main uses of a water storage tank is to maintain a water supply at constant hydrodynamic voltage.

Figure (17) is a schematic diagram of a typical town water supply. Water is pumped from the reservoir up into the water tank where a constant height h and therefore constant voltage  $\rho gh$  is maintained. The houses in the town all draw constant voltage water from this tank.

Let us see what would happen if the water tank was too small. As soon as several houses started using water, the level h in the tank would drop and the pump at the reservoir would have to come on. The pump would raise the level back to h and shut off. Then the level would drop again and the pump would come on again. The result would be that the hydrodynamic voltage or water pressure supplied to the town would vary and customers might complain.

On the other hand if the town water tank has a large cross-sectional area and therefore large capacitance C, a few houses drawing water would have very little effect on the level h and therefore voltage  $\rho$ gh of the water. The town would have a constant voltage water supply and the water company could pump water from the reservoir at night when electricity rates were low.

We will see that one of the important uses of electrical capacitors in electric circuits is to maintain constant or nearly constant electric voltages. There is an accurate analogy to the way the town water tank maintains constant voltage water. If we use too small a capacitor, the electrical voltage will also fluctuate when current is drawn.



#### Figure 17

Town water supply. By maintaining a constant height h of water in the storage tank, all the water supplied to the town has a constant hydrodynamic voltage  $V_h \rho gh$ .

## **Electrical Capacitance**

Figure (18) is a repeat of the sketches of the parallel plate capacitor discussed in Chapter 26. The important features of the capacitor are the following. We have two metal plates of area A separated by a distance d. The positive plate shown on top has a charge + Q, the bottom plate a charge – Q. Since the area of the plates is A, the surface charge density on the inside of the plate is

$$\sigma = \frac{Q}{A} \tag{27}$$

In Chapter 26, page 26-3, we saw that a charge density  $\sigma$  on the surface of a conductor produced an electric field of strength

$$E = \frac{\sigma}{\varepsilon_0}$$
(28)

perpendicularly out of the conductor. In Figure (18) this field starts at the positive charge on the inside of the upper plate and stops at the negative charge on the inside surface of the bottom plate.

Recall that one form of electric voltage is the electric potential energy of a unit test charge. To lift a positive unit test charge from the bottom plate to the top one requires an amount of work equal to the force E on a unit charge times the distance d the charge was lifted. This work E \* d is equal to the increase of the potential energy of the unit charge, and therefore to the increase in voltage in going from the bottom to the top plate. If we say that the bottom plate is at a voltage V = 0, then the voltage at the top plate is

$$\mathbf{V} = \mathbf{E} \, \mathbf{d} \tag{29}$$



#### Figure 18

The parallel plate capacitor. If we place charges + Q and – Q on plates of area A, the charge density on the plates will be  $\sigma = Q/A$ , the electric field will be  $E = \sigma/\epsilon_0$  and the voltage between the plates V = Ed. Using Equation (28) for E and Equation (27) for  $\sigma$ , we get the relationship

$$V = \frac{\sigma}{\varepsilon_0} d = \frac{Q}{\varepsilon_0 A} d$$

or

$$Q = \left(\frac{\varepsilon_0 A}{d}\right) V$$
(30)

which is our old Equation (26-10).

As in our hydrodynamic analogy, we see that the quantity of charge Q stored in the capacitor is proportional to the voltage V on the capacitor. Again we call the proportionality constant the capacitance C

$$Q = CV \qquad \begin{array}{c} definition of \\ electrical \\ capacitance \end{array} \tag{31}$$

Comparing Equations (30) and (31) we see that the formula for the capacitance C of a parallel plate capacitor is

$$C = \frac{\varepsilon_0 A}{d}$$

$$\begin{array}{c} capacitance of \ a \\ parallel \ plate \\ capacitor \ of \ area \ A, \\ plate \ separation \ d \end{array} \tag{32}$$

For both the parallel plate capacitor and the cylindrical water tank, the capacitance is proportional to the crosssectional area A. The new feature for the electrical capacitor is that the capacitance increases as we make the plate separation d smaller and smaller.

Our parallel plate capacitor is but one example of many kinds of capacitors used in electronic circuits. In some, the geometry of the metal conductors is different, and in others the space between the conductors is filled with a material called a dielectric which increases the effective capacitance. But in all common capacitors the amount of charge Q is proportional to voltage V across the capacitor, i.e. Q=CV, where C is constant independent of the voltage V and in most cases independent of the temperature. The dimensions of capacitance C are coulombs per volt, which is given the name **farad** in honor of Michael Faraday who pioneered the concept of an electric field. Although such an honor may be deserved, this is one more example of the excessive use of names in the MKS system that make it hard to follow the dimensions in a calculation.

To get a feeling for the size of a farad, suppose that we have two metal plates with an area  $A = 0.1 \text{ meter}^2$  and make a separation d = 1 millimeter =  $10^{-3}$  meters. These plates will have a capacitance C given by

$$C = \frac{\varepsilon_0 A}{d} = \frac{9 \times 10^{-12} \times .1}{10^{-3}}$$
$$= 9 \times 10^{-10} \text{ farads}$$

which is about one billionth of a farad. If you keep the separation at 1 millimeter you would need plates with an area of 100 million square meters (an area 10 kilometers on a side) to have a capacitance of 1 farad.

Commercial capacitors used in electronic circuits come in various shapes like those shown in Figure (19), and in an enormous range of values from a few farads down to  $10^{-14}$  farads.

Our calculation of the capacitance of a parallel plate capacitor demonstrates that it is not an easy trick to produce capacitors with a capacitance of  $10^{-6}$  farads or larger. One technique is to take two long strips of metal foil separated by an insulator, and roll them up into a small cylinder. This gives us a large plate area with a reasonably small separation, stuffed into a relatively small volume.

In a special kind of a capacitor called an electrolytic capacitor, the effective plate separation d is reduced to almost atomic dimensions. Only this way are we able to create the physically small 1 farad capacitor shown in Figure (19). The problem with electrolytic capacitors is that one side has to be positive and the other negative, as marked on the capacitor. If you reverse the voltage on an electrolytic capacitor, it will not work and may explode.

## **Exercise 4 - Electrolytic Capacitor**

In an *electrolytic capacitor*, one of the plates is a thin aluminum sheet and the other is a conducting dielectric liquid surrounding the aluminum. A nonconducting oxide layer forms on the surface of the aluminum and plays the same role as the air gap in the parallel plate capacitors we have been discussing. The fact that the oxide layer is very thin means that you can construct a capacitor with a very large capacitance in a small container.



For this problem, assume that you have a dielectric capacitor whose total capacitance is 1 farad, and that the oxide layer acts like an air gap  $10^{-7}$  meters thick in a parallel plate capacitor. From this, estimate the area of the aluminum surface in the capacitor.



#### Figure 19

Examples of capacitors used in electronic circuits. The one on the right is a variable capacitor whose plate area is changed by turning the knob. The square black capacitor is a 4 farad electrolytic. Its capacitance is one million times greater than the tall regular capacitor behind it.

# **ENERGY STORAGE IN CAPACITORS**

In physics, one of the important uses of capacitors is energy storage. The advantage of using capacitors is that large quantities of energy can be released in a very short time. For example, Figure (20) is a photograph of the Nova laser at the Lawrence Livermore National Laboratory. This laser produces short, but very high energy pulses of light for fusion research. The laser is powered by a bank of capacitors which, for the short length of time needed, can supply power at a rate about 200 times the power generating capacity of the United States.

The easiest way to determine the amount of energy stored in a capacitor is to calculate how much work is required to charge up the capacitor. In Figure (21) we have a capacitor of capacitance C that already has a charge + Q on the positive plate and -Q on the negative plate. The voltage V across the capacitor is related to Q by Equation (12), Q = CV.

Now let us take a charge dQ out of the bottom plate, leaving a charge -(Q + dQ) behind, and lift it to the top plate, leaving (Q + dQ) there. The work dW we do to lift the charge is equal to dQ times the work required to lift a unit test charge, namely dQ times the voltage V

$$dW = VdQ$$

or replacing V by Q/C, we have

$$dW = \frac{Q}{C}dQ$$
(33)



#### Figure 21





The Nova laser, powered by a bank of capacitors. While the laser is being fired, the capacitors supply 200 times as much power as the generating capacity of the United States.

You can see from Equation (33) that when the capacitor is uncharged and we lift the first dQ, no work is required because there is no field yet in the capacitor. However once there is a big charge on the capacitor, much work is required to lift an additional dQ. The total amount of work to charge the capacitor from zero charge to a final charge  $Q_f$  is clearly given by the integral

Work = 
$$\int dW = \int_0^{Q_f} \frac{Q}{C} dQ$$

The fact that the capacitance C is a constant, means that we can take it outside the integral and we get

$$W = \frac{1}{C} \int_{0}^{Q_{f}} Q dQ = \frac{Q_{f}^{2}}{2C}$$
(34)

Since it is easier to measure the final voltage V rather than the charge  $Q_f$  in a capacitor, we use  $Q_f = CV$  to rewrite Equation (34) in the form

$$\frac{\text{Energy stored}}{\text{in a capacitor}} = \frac{CV_f^2}{2}$$
(35)

The energy stored is proportional to the capacitance C of the capacitor, and the square of the voltage V.

## **Energy Density in an Electric Field**

Equation 35 can be written in a form that shows that the energy stored in a capacitor is proportional to the square of the strength of the electric field. Substituting  $V_f = E \times d$  and  $C = \epsilon_0 A/d$  into Equation 35 gives

Energy stored  
in a capacitor 
$$= \frac{C V_f^2}{2} = \frac{1}{2} \frac{\varepsilon_0 A}{d} \times E^2 d^2$$
  
 $= \frac{\varepsilon_0 E^2}{2} \times A d = \frac{\varepsilon_0 E^2}{2} \times \begin{pmatrix} Volume \\ Inside \\ capacitor \end{pmatrix}$ 

where we note that  $A \times d$  is the volume inside the capacitor.

Since the energy stored in the capacitor is proportional to the volume occupied by the electric field, we see that the energy per unit volume, the energy density, is simply given by

$$\frac{\text{Energy}}{\text{density}} = \frac{\varepsilon_0 E^2}{2}$$
(36)

This result, that the energy density in an electric field is proportional to the square of the strength of the field, turns out to be a far more general result than we might expect from the above derivation. It applies not just to the uniform electric field in an idealized capacitor, but to electric fields of arbitrary shape.

## **Exercise 5**

A parallel plate capacitor consists of two circular aluminum plates with a radius of 11 cm separated by a distance of 1 millimeter. The capacitor is charged to a voltage of 5 volts.

a) What is the capacitance, in farads, of the capacitor?

b) Using Equation 35, calculate the energy stored in the capacitor.

c) What is the magnitude of the electric field E between the plates?

d) Using equation 36, calculate the energy density in the electric field.

e) What is the volume of space, in cubic meters, between the plates?

f) From your answers to parts d) and e), calculate the total energy in the electric field between the plates.Compare your answer with your answer to part b-

g) Using Einstein's formula  $E = mc^2$ , calculate the mass, in kilograms, of the electric field between the plates.

h) The mass of the electric field is equal to the mass of how many electrons?

# CAPACITORS AS CIRCUIT ELEMENTS

Figure (22) is a simple circuit consisting of a battery of voltage  $V_b$  and a capacitor of capacitance C. The standard circuit symbol for a capacitor is -||----, which is a sketch of a parallel plate capacitor.

When the battery is attached to the capacitor, the upper plate becomes positively charged and the lower one negatively charged as shown. The upper plate could actually become positively charged either by positive charge flowing into it or negative charge flowing out it does not matter. We have followed our convention of always showing the direction of positive currents, thus we show i flowing into the positive plate and out of the negative one.

We have also followed our convention of labeling the voltage rises with an arrow pointing in the direction of the higher voltage. The voltage  $V_c$  on the capacitor is related to the charge Q stored by the definition of capacitance,  $V_c = Q/C$ .



#### Figure 22

A battery and a capacitor in a circuit. We have drawn the diagram showing positive current flowing into the top plate and out of the bottom plate. The upper plate could have become positively charged by having a negative current flowing out of it. The arrow designating the voltage on the capacitor points in the direction of the voltage rise.

Applying Kirchoff's law to Figure (22), i.e., setting the sum of the voltage rises around the circuit equal to zero, we get

$$V_{b} + (-V_{c}) = V_{b} - Q/C = 0$$

$$Q = CV_{b}$$
(37)

Thus we get a relatively straightforward result for the amount of charge stored by the battery.

For something a little more challenging, we have connected two capacitors in parallel to a battery as shown in Figure (23). Because single wires go all the way across the top and across the bottom, the three voltages  $V_b$ ,  $V_1$  and  $V_2$  must all be equal, and we get

$$Q_1 = C_1 V_b \qquad \qquad Q_2 = C_2 V_b$$

The total charge Q stored on the two capacitors in parallel is therefore

$$Q = Q_1 + Q_2 = (C_1 + C_2)V_b$$

Comparing this with Equation (37), we see that two capacitors in parallel store the same charge as a single capacitor C given by

$$C = C_1 + C_2 \qquad \begin{array}{c} capacitors \\ attached in \\ parallel \end{array} \tag{38}$$

Comparing this result with Equation (16), we find that for capacitors in parallel or resistors in series, the effective capacitance or resistance is just the sum of the values of the individual components.



#### Figure 23

Capacitors connected in parallel. The three voltages  $V_b$ ,  $V_1$  and  $V_2$  must all be level because the wires go all the way across the three elements.

In Figure (24) we have two capacitors in series. The trick here is to note that all the charge that flowed out of the bottom plate of  $C_1$  flowed into the top plate of  $C_2$ , as indicated in the diagram. But if there is a charge – Q on the bottom plate of  $C_1$ , there must be an equal and opposite charge + Q on the top and we have  $Q_1 = Q$ . Similarly we must have  $Q_2 = Q$ .

To apply Kirchoff's law, we set the sum of the voltage rises to zero to get

$$V_{b} + \frac{-Q_{1}}{C_{1}} + \frac{-Q_{2}}{C_{2}} = 0$$

Setting  $Q_1 = Q_2 = Q$  gives

$$V_{\rm b} = Q\left(\frac{1}{C_1} + \frac{1}{C_2}\right) \tag{39}$$

Comparing Equation (39) with Equation (37) in the form  $V_b = Q/C$  we see that

$$\frac{1}{C} = \frac{1}{C_1} + \frac{1}{C_2}$$

$$\begin{array}{c} capacitors \\ attached in \\ series \end{array} (40)$$

is the formula for the effective capacitance of capacitors connected in series. This is analogous to the formula for parallel resistors.



#### Figure 24

Capacitors in series. In this case the sum of  $V_1$ and  $V_2$  must be equal to the battery voltage  $V_b$ .

It is interesting to note that for storing charge, parallel capacitors are more efficient because the charge can flow into both capacitors as seen in Figure (23). When the capacitors are in series, charge flowing out of the bottom of one capacitor flows into the top of the next, and we get no enhancement in charge storage capability. What we do get from series capacitors is higher voltages, the total voltage rise across the pair is the sum of the voltage rise on each.

#### **Exercise 6**

You have a 5 microfarad (abbreviated  $5\mu$ f) capacitor and a 10  $\mu$ f capacitor. What are all the values of capacitor you can make from these two?



#### Figure 25

An RC circuit. When the mercury switch is closed, the capacitor quickly charges up to a voltage  $V_C = V_b$ . When the switch is opened, the capacitor discharges through the resistor.



Figure 26

Capacitor discharge. When the switch is open, the only part of the circuit we have to look at is the capacitor discharging through the resistor.

# THE RC CIRCUIT

The capacitor circuits we have discussed so far are not too exciting. When you are working with an electronic circuit you do not hitch capacitors together in series or parallel, you simply go to the parts drawer and select a capacitor of the desired value.

If we add a resistor to the circuit as shown in Figure (25), we begin to get some interesting results. The circuit is designed so that if the mercury switch is closed, the capacitor is charged up to a voltage  $V_b$  by the battery. Then, at a time we will call t = 0, the switch is opened, so that the capacitor will discharge through the resistor. During the discharge, the battery is disconnected and the only part of the circuit that is active is that shown in Figure (26). (The reason for using a mercury switch was to get a clean break in the current. Mechanical switches do not work well.)

Figure (27) shows the capacitor voltage just before and for a while after the switch was opened. We are looking at the experimental results of discharging a  $C = 10^{-6}$  farad (one microfarad) capacitor through an  $R = 10^4$  ohm resistor. We see that a good fraction of the capacitor voltage has decayed in about 10 milliseconds ( $10^{-2}$  seconds).

To analyze the capacitor discharge, we apply Kirchoff's law to the circuit in Figure (26). Setting the sum of the voltage rises around the circuit equal to zero gives

$$V_{\rm C} - V_{\rm R} = 0$$
$$\frac{Q}{C} - iR = 0$$

which can be written in the form

$$-i + \frac{Q}{RC} = 0 \tag{41}$$

The problem with Equation (41) is that we have two unknowns, i and Q, and only one equation. We need to find another relationship between these variables in order to predict the behavior of the circuit. The additional relationship is obtained by noting that the current i is the number of coulombs per second flowing out of the capacitor. In a short time dt, an amount of charge dQ that leaves the capacitor is given by

$$dQ = idt$$
 (42)

Dividing Equation (42) through by dt, and including a minus sign to represent the fact that i is causing a decrease in the charge Q in the capacitor, we get

$$\left| \begin{array}{c} \frac{\mathrm{d}Q}{\mathrm{d}t} = -\mathrm{i} \\ a \ capacitor \end{array} \right| \qquad (43)$$

Substituting Equation (43) in (41) gives one equation for the unknown Q

$$\frac{dQ}{dt} + \frac{Q}{RC} = 0$$
equation for  
the discharge  
of a capacitor
(44)

## **Exponential Decay**

The next problem is that Equation (44) is a differential equation, of a type we have not yet discussed in the text. We met another kind of differential equation in our discussion of harmonic motion, a differential equation that involved second derivatives and had oscillating sinusoidal solutions. Equation (44) has only a first derivative, and produces a different kind of solution. There are two principle ways of solving a differential equation. One is to use a computer, and the other, the so-called analytic method, is to guess the answer and then check to see if you have made the correct guess. We will first apply the analytic method to Equation (44). In the supplement we will show how a computer solution is obtained.

The important thing to remember about a differential equation is that the solution is a shape or a curve, not a number. The equation  $x^2 = 4$  has the solutions  $x = \pm 2$ , the solution to Equation (44) is the curve shown in Figure (27). One of the advantages of working with electric circuits is that the theory gives you the differential equation, and the equipment in the lab allows you to look at the solution. The curve in Figure (27) is the voltage on the capacitor recorded by the computer based oscilloscope we used to record the motion of air carts and do the analysis of sound waves. We are now using the device as a voltmeter that draws a picture of the voltage.

The curve in Figure (27) is well known to scientists in many fields as an *exponential decay*. Exponential decays are best known in studies of radioactive decay and are associated with the familiar concept of a half life. Let us first write down the formula for an exponential decay, check that the formula is, in fact, a solution to Equation (44) and then discuss the special properties of the curve.



Figure 27 Experimental results from discharging the one microfarad  $(10^{-6}f)$  capacitor through a 10 k ohm  $(10^4 \Omega)$  resistor. The switch, shown in Figure 26, is thrown at time t = 0.

If Figure (27) represents an exponential decay of the capacitor voltage  $V_c$ , then  $V_c$  must be of the form

$$V_{\rm C} = V_0 e^{-\alpha t} \tag{45}$$

where  $V_0$  and  $\alpha$  are constants to be determined. Since Equation (44) is in terms of Q rather than  $V_C$ , we can use the definition of capacitance Q = CV to rewrite Equation (45) as

$$CV_{C} = CV_{0}e^{-\alpha t}$$

$$Q = Q_{0}e^{-\alpha t}$$
(46)

where  $Q_0 = CV_0$ .

Differentiating Equation (46) with respect to time gives

$$\frac{\mathrm{d}Q}{\mathrm{d}t} = -\alpha Q_0 e^{-\alpha t} = -\alpha Q \tag{47}$$

This result illustrates one of the properties of an exponential decay, namely that the derivative of the function is proportional to the function itself (here  $dQ/dt = -\alpha Q$ ).

Substituting Equation (47) into our differential Equation (44) gives

$$\frac{\mathrm{dQ}}{\mathrm{dt}} + \frac{\mathrm{Q}}{\mathrm{RC}} = 0 \tag{44}$$

$$-\alpha Q + \frac{Q}{RC} = 0 \tag{48}$$

The Qs cancel in Equation (48) and we get

$$\alpha = \frac{1}{\text{RC}} \tag{49}$$

Thus the coefficient of the exponent is determined by the differential equation.

## Exercise 7

Determine the constant V<sub>0</sub> in Equation (45) from Figure (27), by noting that at t = 0,  $e^{-\alpha t} = e^{0} = 1$ .

## The Time Constant RC

Substituting Equation (49) for  $\alpha$  back into our formula for V<sub>C</sub> gives

$$V_{\rm C} = V_0 e^{-(t/RC)} \tag{50}$$

Since the exponent (t/RC) must be dimensionless, the quantity RC in the denominator must have the dimensions of time. Since R is in ohms and C in farads, we must have

ohms 
$$*$$
 farads = seconds (51)

We have mentioned that units in electrical calculations are hard to follow, and this is a prime example. We leave it as a challenge to go back and actually show, from the definition of the ohm and of the farad, that the product ohms times farads comes out in seconds.

The quantity RC that appears in Equation (50) is known as the *time constant* for the decay. At the time t = RC, the voltage V<sub>C</sub> has the value

$$V_{C} (at t = RC) = V_{b}e^{-RC/RC}$$
  
=  $V_{b}e^{-1} = \frac{V_{b}}{e}$  (52)

I.e., in one time constant RC, the voltage has decayed to 1/e = 1/2.7 of its initial value.

To see if this analysis works experimentally, we have gone back to Figure (27) and marked the time RC. In that experiment

$$R = 10^4 \text{ ohms}$$
$$C = 10^{-6} \text{ farads}$$

thus

$$RC = 10^{-2} \text{ ohm farads}$$
  
= 10<sup>-2</sup> seconds (53)  
= 10 milliseconds

We see that at a time T = RC, the voltage dropped from  $V_b = 4$  volts to V(t=RC) = 1.5 volts, which is down by a factor 1/e = 1/2.7.

If we wait another time constant, until t = 2RC, we have

$$V_{\rm C}$$
 (at t = 2RC) =  $V_{\rm b}e^{-2} = \frac{V_{\rm b}}{e^2}$ 

and we get another factor of e = 2.7 in the denominator. In Figure (28) the voltage is down to 4/(2.7\*2.7) = .55 volts at t = 2RC. After each succeeding time constant **RC**, the voltage drops by another factor of 1/e.

## **Half-Lives**

When you first studied radioactive decay, you learned about half-lives. A half-life was the time it took for half of the remaining radioactive particles to decay. Wait another half-life and half of those are gone. In our description of the exponential decay, the time constant RC is similar to a half-life, but just a bit longer. When we wait for a time constant, the voltage decays down to 1/2.7 of its initial value rather than 1/2 of its initial value. In Figure (29) we compare the half-life t 1/2 and the time constant RC. Although the half-life is easier to explain, we will see that the time constant RC provides a more convenient unit of time for the analysis of the exponential decay curve.



Exponential decay of the voltage in the capacitor. In a time t = RC the voltage drops by a factor 1/e = 1/2.7. In the next time interval RC, the voltage drops by another factor of 1/e.



#### Figure 29

The time it takes the voltage to drop to half its initial value, what we could call the "half life" of the voltage, is a bit shorter than the time constant RC.

## **Initial Slope**

One of the special features of a time constant is the fact that the initial slope of the curve intercepts the zero value one time constant later, as illustrated in Figure (30). It does not matter where we take our initial time to be. Pick any point on the curve, draw a tangent line at that point, and the tangent line intercepts the V = 0line one time constant RC later. This turns out to be the most convenient way to determine the time constant from an experimental curve. Try it yourself in the following exercise.



The initial slope of the discharge curve intersects the  $V_C = 0$  origin at a time t = RC, one time constant later. This fact provides an easy way to estimate the time constant for an exponential curve.

#### **Exercise 8**

In Figure (31) we have the experimental voltage decay for an RC circuit where R is known to be  $10^5$  ohms. Determine the time constant for this curve and from that find out what the value of the capacitance C must have been.

## **The Exponential Rise**

Figure (32) is a circuit in which we use a battery to charge a capacitor through a resistance R. The experimental result, for our capacitor  $C = 10^{-6}$  farads,  $R = 10^4$  ohms is shown in Figure (33). Here the capacitor starts charging relatively fast, then the rate of charging slows until the capacitor voltage finally reaches the battery voltage V<sub>b</sub>.

If the shape of the curve in Figure (33) looks vaguely familiar, it should. Turn the curve over and it looks like our exponential decay curve with time increasing toward the left. If that is true, then the initial slope of the charge up curve should intercept the  $V_C = V_b$  line one time constant later as shown in Figure (34), and it does.

Because we have a battery in the circuit while the capacitor is charging up, the analysis is a bit more messy than for the capacitor discharge. (You should be able to repeat the analysis of the capacitor discharge on your own, and at least be able to follow the steps for analyzing the charge up.)



**Figure 31** Experimental results for the discharge of a capacitor through a  $10^5 \Omega$  resistor.

In Figure (32) we have drawn the circuit diagram and indicated the voltage rises around the circuit. Kirchoff's law there gives

$$V_{b} + (-V_{R}) + (-V_{C}) = 0$$
  
 $V_{b} - iR - \frac{Q}{C} = 0$  (54)

One difference between Figure (32) for the charge up and Figure (27) for the discharge, is that for the charge up, the current i is flowing into, rather than out of, the capacitor. Therefore in a time dt the charge in the capacitor **increases** by an amount dQ = idt, and we have

$$\frac{\mathrm{d}Q}{\mathrm{d}t} = +\mathrm{i} \tag{55}$$

Using Equation (55) in (54), dividing through by R, and rearranging a bit, gives

$$\frac{\mathrm{dQ}}{\mathrm{dt}} + \frac{\mathrm{Q}}{\mathrm{RC}} = \frac{\mathrm{V}_{\mathrm{b}}}{\mathrm{R}} \tag{56}$$

It is the term on the right that makes this differential equation harder to solve. A simple guess like the one we made in Equation (46) does not work, and we have to try a more complicated guess like

$$Q = A + Be^{-\alpha t}$$
(57)

When you take a course in solving differential equations, much of the time is spent learning how to guess the form of solutions. For now, let us just see if the guess in Equation (57) can be made to work for some value of the constants A, B, and  $\alpha$ .



**Figure 32** *Charging up a capacitor through a resistor.* 

Differentiating Equation (57) with respect to time gives

$$\frac{\mathrm{d}Q}{\mathrm{d}t} = -\alpha \mathrm{B}\mathrm{e}^{-\alpha t} \tag{58}$$

Substituting Equations (57) and (58) into (56) gives

$$-\alpha B e^{-\alpha t} + \frac{A}{RC} + \frac{B}{RC} e^{-\alpha t} = \frac{V_b}{R}$$
(59)

The only way we can satisfy Equation (59) is have the two terms with an  $e^{-\alpha t}$  cancel each other. I.e., we must have

$$-\alpha B + \frac{B}{RC} = 0$$

$$\alpha = \frac{1}{RC}$$
(60)

which is a familiar result.

The remaining terms in Equation (59) give

$$\frac{A}{RC} = \frac{V_b}{R} \qquad A = CV_b \tag{61}$$

Putting the values for A and  $\alpha$  (Equations 61 and 60) back into our guess (Equation 57), we get

$$Q = CV_b + Be^{-t/RC}$$
(62)

The final step is to note that at time t = 0, Q = 0, so that

$$0 = CV_b + Be^0 \qquad B = -CV_b$$

thus our final result is

$$Q = CV_{b} \left( 1 - e^{-t/RC} \right)$$
(63a)

We can express this result in terms of voltages if we divide through by C and use 
$$V_C = Q/C$$
, we get

$$V_{\rm C} = V_{\rm b} \left( 1 - e^{-t/{\rm RC}} \right) \qquad \begin{array}{c} charging \ a \\ capacitor \end{array} \tag{63}$$

(We warned you that the addition of just one more term to our differential equation would make it messier to solve.)

The answer, Equation (63) is the standard form for an exponential rise. It is in fact just our exponential decay curve turned upside down, and Figure (34) represents the easy way to determine the time constant RC from experimental data.



#### Figure 34

If you continue along the initial slope line of the charge-up curve, you intersect the final voltage  $V_b$  at a time t = RC, one time constant later. Turn this diagram upside down and it looks like Figure 31.



#### Figure 33

Plot of the capacitor voltage versus time for the charging up of a capacitor through a resistor. If you turn the diagram upside down, you get the curve for the discharge of a capacitor.

#### **Exercise 9**

Figure (35) shows the voltage across a capacitor C being charged through a resistance R. Given that  $C = 1.0 \times 10^{-8}$  farads, estimate the value of R.



#### Figure 35

Given the experimental results for the charge-up of a capacitor, determine the value of the resistance R. (Answer R = 68K)



#### Figure 36

A neon bulb. When the voltage across the wires reaches a threshold value, typically around 100 volts, the neon gas starts to glow, and the gas suddenly changes from an insulator to a conductor.

# THE NEON BULB OSCILLATOR

We will end this chapter on basic electric circuits with a discussion of an electronic device called a neon bulb oscillator. This is conceptually the world's simplest electronic device that does something useful—it oscillates and its frequency of oscillation can be adjusted. The device is not practical, for it is hard to adjust, its waveform is far from being a pure sinusoidal shape, and it requires a relatively high voltage power supply. But when you work with this apparatus, you will begin to get a feeling for the kind of tricks we pull in order to make useful apparatus.

## The Neon Bulb

The new circuit element we will add to our neon oscillator circuit is the common neon bulb which glows orange and is often used as a night light. The bulb, which we will designate by the symbol -(•)— is simply a small glass tube with neon gas inside and two wires as shown in Figure (36). The bulb turns on when there is a large enough voltage difference between the wires that the neon gas becomes ionized and starts to glow. For typical neon bulbs, the glow starts when the voltage reaches approximately 100 volts. When the bulb is glowing the neon gas is a good conductor and the bulb is like a closed switch. When the bulb is not glowing, the gas is inert and the bulb is like an open switch.

A given neon bulb has a rather consistent voltage  $V_f$  (*firing voltage*) at which it turns on, and voltage  $V_q$  (*quenching voltage*) at which it shuts off. In a typical bulb  $V_f$  may be 100 volts and  $V_q$  equal to 40 volts. These numbers will, however, vary from bulb to bulb. When a neon bulb is included in a circuit, it acts like an automatic switch, closing (turning on) when the voltage across it reaches  $V_f$  and opening (shutting off) when the voltage drops to  $V_q$ .

## **The Neon Oscillator Circuit**

We can make a neon oscillator using the circuit shown in Figure (37). The left hand part of the circuit is just the RC circuit we used in Figure (32) to charge up a capacitor. The only really new feature is the neon bulb in parallel with the capacitor. The recording voltmeter, indicated by the symbol  $-\sqrt{V}$  is there to record the capacitor voltage V<sub>C</sub>.

The output of the neon oscillator circuit is shown in Figure (38). Initially the capacitor is charging up just as it did in Figure (33). During the charge up, the neon bulb is off; it is like an open switch and might as well not be there. The effective circuit is shown in Figure (39). When the voltage on the capacitor (and on the neon bulb) reaches the firing voltage  $V_{f}$ , the neon bulb turns on and acts like a short circuit as shown in Figure (40). It is not exactly a short circuit, the neon bulb and the wire leads have some small resistance. But the resistance is so small that the capacitor rapidly discharges through the bulb, and the capacitor voltage drops almost instantly.

When the capacitor and neon bulb voltage  $V_c$  drops to the bulb quenching voltage  $V_q$ , the bulb shuts off, and the capacitor starts charging up again. As seen in Figure (38), this process keeps repeating and we get the oscillating voltage shown.

For the last cycle in Figure (38), we opened a switch to disconnect the neon bulb, allowing the capacitor to charge up all the way to the power supply voltage  $V_b$ . This allowed us to display all three voltages  $V_b$ ,  $V_f$ , and  $V_q$  on one experimental plot.





**Figure 37** *Neon bulb oscillator circuit.* 



Figure 39

Effective circuit while the neon bulb is off.



**Figure 40** *Effective circuit while the neon bulb is glowing.* 

#### Figure 38

Experimental output of a neon oscillator circuit. The capacitor charges up until the voltage reaches the neon bulb firing voltage  $V_f$ , at which point the neon bulb turns on and the voltage rapidly drops. When the voltage has fallen to the quench voltage  $V_q$ , the neon bulb shuts off and the capacitor voltage starts to rise again. On the last cycle, we opened a switch to disconnect the neon bulb, allowing the capacitor to charge up all the way to the power supply voltage  $V_b$ .

(A voltage divider was used to measure these high voltages. In the figure, the voltage scale has been corrected to represent the actual voltage on the capacitor.)

## **Period of Oscillation**

To calculate the period of oscillation, we start with the diagram of Figure (41) showing a cycle of the oscillation superimposed upon the complete charge up curve which starts at  $V_C = 0$  and goes up to  $V_C = V_b$ . This curve is given by the formula

$$V_{\rm C} = V_{\rm b} \left( 1 - e^{-t/R\rm C} \right) \tag{63}$$

What we want to calculate is the time  $T = (t_2 - t_1)$  it takes for the capacitor to charge up from a voltage  $V_q$  to  $V_f$ .

At time  $t_1$ ,  $V_C = V_q$  and Equation (63) gives

$$V_{q} = V_{b} \left( 1 - e^{-t_{1}/RC} \right) \tag{64}$$

At time  $t = t_2$ ,  $V_C = V_f$  and we have

$$V_{\rm f} = V_{\rm b} \left( 1 - e^{-t_2/{\rm RC}} \right)$$
 (65)

Equation (64) and (65) can be rearranged to give

$$e^{-t_1/RC} = 1 - \frac{V_q}{V_b}$$
 (66)

$$e^{-t_2/RC} = 1 - \frac{V_f}{V_b}$$
 (67)



#### Figure 41

Determining the period T of the oscillation. The formulas for  $t_1$  and  $t_2$  are obtained from the capacitor charge up equations

$$V_f = V_b (1 - e^{-t_1/RC})$$
$$V_q = V_b (1 - e^{-t_2/RC})$$

The messy part is extracting the period  $T = t_2 - t_1$ from these equations. Dividing Equation (66) by Equation (67) gives

$$e^{(t_2 - t_1)/RC} = \frac{1 - V_q/V_b}{1 - V_f/V_b} = \frac{V_b - V_q}{V_b - V_f}$$
(68)

where we used the fact that

$$\frac{e^{-\alpha}}{e^{-\beta}} = e^{\beta - \alpha}$$

Taking the logarithm of Equation (68), using  $\ln(e^{\alpha}) = \alpha$ , we have

$$\frac{t_2 - t_1}{RC} = \ln\left(\frac{V_b - V_q}{V_b - V_f}\right)$$

or using  $T = t_2 - t_1$  we get for the period of oscillation

$$T = (RC) ln \left( \frac{V_b - V_q}{V_b - V_f} \right) \qquad period of \\ neon oscillator \qquad (69)$$

Equation (69) was a bit messy to derive, and it is not a fundamental result that you need to memorize. You are unlikely to meet a neon oscillator except in an introductory physics lab. But we have used the theory developed in this chapter to make an explicit prediction that can be tested in the laboratory.

#### Exercise 10

See how well Equation (69) applies to the experimental data of Figure (38). (The marked values on resistors and capacitors are usually accurate only to within  $\pm$  10%.)
Equation (69) provides clear instructions on how to change the period or frequency of a neon oscillator. The easiest way to make major changes in the period is to change the time constant RC. Because of the ease with which we can select different values of R and C (typical values of R ranging from  $10^2$  ohms to  $10^8$  ohms, and typical values of C from  $10^{-4}$  to  $10^{-12}$  farads), a large range of time constants RC are available. However high frequencies are limited by the characteristics of the neon bulb. We found it difficult to get the circuit to oscillate faster than 30 cycles per second.

Adjusting the battery voltage  $V_b$  changes the shape of the neon oscillator wave and also allows fine adjustments in the period .

## **Experimental Setup**

An experimental problem you face while working with the neon oscillator circuit, is that the voltages of interest range up to 100 volts or more. Modern oscilloscopes or recording voltmeters tend to operate in the range of +5 to -5 volts or less, and should not be attached to a voltage source of the order of 100 volts. This problem can be solved by using the voltage divider circuit discussed in Exercise (3), and shown in Figure (42).

For a standard laboratory experiment, we have found it convenient to mount, in one box, the voltage divider, neon bulb, and switch - the components shown inside the dotted rectangle of Figure (42). This reduces student exposure to high voltages and guarantees that the voltmeter will be exposed to voltages 1000 times smaller than those across the capacitor.

In Figure (43), we have recorded the entire voltage range of the experiment, starting from an uncharged capacitor. We first opened the switch above the neon bulb in Figure (42), and let the capacitor charge up to the full voltage  $V_b$ . Then closing the switch allowed the capacitor voltage to oscillate between  $V_f$  and  $V_q$  as seen in Figure (38). While the actual capacitor voltage ranges from 0 to 100 volts, the recording voltmeter shows a range of 0 to 100 millivolts because of the 1000 to 1 voltage divider.



# Figure 42

Neon oscillator circuit with voltage divider. The switch above the neon bulb allows us to disconnect the bulb from the circuit. It is convenient to mount, in a single box, the components within the dotted rectangle.



Full range of voltages from the neon oscillator circuit. The voltage scale is in millivolts because of the voltage divider.

#### **Exercise 11 Review Problem**

Figure (44a) shows the circuit used to observe the discharge of a capacitor. The capacitor is made from the two circular aluminum plates shown in Figure (44b). The plates have a diameter of 22 cm and are separated a distance (d) by small pieces of glass. In Figure (44c), we are observing the discharge of the capacitor through a 10k $\Omega(10^4\Omega)$  resistor. For this discharge, what is the separation (d) of the plates?



**Figure 44a** *Circuit for observing the discharge of a capacitor.* 



**Figure 44b** *The capacitor plates.* 



Voltage during discharge.

# Chapter 28 Magnetism

In our discussion of the four basic interactions, we saw that electric forces are very strong but in most circumstances tend to cancel. The strength of the forces are so great, but the cancellation is so nearly complete that the slightest imbalance in the cancellation leads to important effects such as molecular forces. As illustrated in Figure (18-6) reproduced here, a positively charged proton brought up to a neutral hydrogen atom experiences a net attractive force because the negative charge in the atom is pulled closer to the proton. This net force is the simplest example of the type of molecular force called a covalent bond.

In this chapter we will study another way that the precise balance between attractive and repulsive electric forces can be upset. So far in our discussion of electrical phenomena such as the flow of currents in wires, the charging of capacitors, etc., we have ignored the effects of special relativity. And we had good reason to. We saw that the conduction electrons in a wire move at utterly nonrelativistic speeds, like two



#### Figure 18-6

The net attraction between a positive charge and a neutral atom is caused by a redistribution of charge in the atom. millimeters per minute. One would not expect phenomena like the Lorentz contraction or time dilation to play any observable role whatever in such electrical phenomena.

But, as we shall see, observable effects do result from the tiny imbalance in electric forces caused by the Lorentz contraction. Since these effects are not describable by Coulomb's law, they are traditionally given another name—magnetism. Magnetism is one of the consequences of requiring that the electrical force law and electric phenomena be consistent with the principle of relativity.

Historically this point of view is backwards. Magnetic effects were known in the time of the ancient Greeks. Hans Christian Oersted first demonstrated the connection between magnetic and electric forces in 1820 and James Clerk Maxwell wrote out a complete theory of electromagnetic phenomena in 1860. Einstein did not discover special relativity until 1905. In fact, Einstein used Maxwell's theory as an important guide in his discovery.

If you follow an historical approach, it appears that special relativity is a consequence of electricity theory, and a large number of physics texts treat it that way. Seldom is there a serious discussion of special relativity until after Maxwell's theory of electricity has been developed. This is considered necessary in order to explain the experiments and arguments that lead to the discovery of the special theory. But as we know today, electricity is one of but several basic forces in nature, and all of them are consistent with special relativity. Einstein's famous theory of gravity called **general relativity** can be viewed as a repair of Newton's theory of gravity to make it consistent with the principle of relativity. (This "repair" produced only minor corrections when applied to our solar system, but has sweeping philosophical implications.) If the principle of relativity underlies the structure of all forces in nature, if all known phenomena are consistent with the principle, then it is not especially necessary to introduce special relativity in the context of its historical origins in electromagnetic theory.

In this chapter we are taking a non-historical point of view. We already know about special relativity (from chapter one), and have just studied Coulomb's electrical force law and some simple applications like the electron gun and basic circuits. We would now like to see if Coulomb's law is consistent with the principle of relativity. In some sense, we would like to do for Coulomb's law of electricity what Einstein did to Newton's law of gravity.

## **Two Garden Peas**

In preparation for our discussion of relativistic effects in electricity theory, let us review a homely example that demonstrates both how strong electric forces actually are, and how complete the cancellation must be for the world to act the way it does. Suppose we had two garden peas, each with a mass of about 2 grams, separated by a distance of 1 meter. Each pea would contain about one mole ( $6 \times 10^{23}$ ) of protons in the atomic nuclei, an equal number of electrons surrounding the nuclei. Thus each pea has a total positive charge +Q in the protons given by

total positive  
charge in a  
garden pea 
$$\Rightarrow 6 \times 10^{23} e$$
$$= 6 \times 10^{23} \times 1.6 \times 10^{-19}$$
$$= 10^{5} coulombs$$
(1)

and there is an equal and opposite amount of negative charge in the electrons.

When two peas are separated by a distance of 1 meter as shown in Figure (1), we can think of there being four pairs of electric forces involved. The positive charge in pea (1) repels the positive charge in pea (2) with a force of magnitude

repulsive force  
between positive  
charge in  
the two peas 
$$= \frac{QQ}{4\pi\epsilon_0 r^2}$$
(2)

which gives rise to one pair of repulsive forces. The negative charges in each pea also repel each other with a force of the same magnitude, giving rise to the second repulsive pair of electric forces. But the positive charge in Pea (1) attracts the negative charge in Pea (2), and the negative charge in Pea (1) attracts the positive charge in Pea (2). This gives us two pairs of attractive forces that precisely cancel the repulsive forces.

Let us put numbers into Equation (2) to see how big these cancelling electric forces are. Equation (2) can be viewed as giving the net force if we removed all the electrons from each garden pea, leaving just the pure positive charge of the protons. The result would be

$$\left|\vec{F}\right| = \frac{Q^2}{4\pi\epsilon_0 r^2}$$
$$= \frac{(10^5 \text{ coulombs})^2}{4\pi \times 9 \times 10^{-12} \times (1)^2}$$
$$= 8.8 \times 10^{19} \text{ newtons}$$
(3)





Electric forces between two garden peas. On pea #1, there is the attractive force between the protons in pea #1 and the electrons in pea #2, and between the electrons in pea #1 and the protons in pea #2. The two repulsive forces are between the electrons in the two peas and the protons in the two peas. The net force is zero. To put this answer in a more recognizable form, note that the weight of one metric ton (1000 kg) of matter is

$$F_{g} (1 \text{ metric ton}) = mg$$
$$= 10^{3} \text{ kg} \times 9.8 \frac{\text{m}}{\text{sec}^{2}}$$
$$= 9.8 \times 10^{3} \text{ newtons}$$

Expressing the force between our two positively charged peas in metric tons we get

repulsive force  
between two  
positive peas  
1 meter apart 
$$= \frac{8.8 \times 10^{19} \text{ newtons}}{9.8 \times 10^{3} \text{ newtons/ton}}$$
$$\approx 10^{15} \text{ tons !}$$
(4)

If we stripped the electrons from two garden peas, and placed them one meter apart, they would repel each other with an electric force of  $10^{15}$  tons!! Yet for two real garden peas, the attractive and repulsive electric force cancel so precisely that the peas can lie next to each other on your dinner plate.

#### **Exercise 1**

Calculate the strength of the gravitational force between the peas. How much stronger is the uncancelled electric force of Equation (3)?

With forces of the order of  $10^{15}$  tons precisely canceling in two garden peas, we can see that even the tiniest imbalance in these forces could lead to striking results. An imbalance of one part in  $10^{15}$ , one part in a million billion, would leave a one ton residual electric force. This is still huge. We have to take seriously imbalances that are thousands of times smaller. One possible source of an imbalance is the Lorentz contraction, as seen in the following thought experiment.

# A THOUGHT EXPERIMENT

In our previous discussion of electric currents, we had difficulty drawing diagrams showing the electrons flowing through the positive charge. To clarify the role of the positive and negative charge, we suggested a model of a copper wire in which we think of the positive and negative charge as being attached to separate rods as shown in Figure (27-5a) repeated here. In that model the rods have equal and opposite charge to represent the fact that the copper wire is electrically neutral, and the negative rod is moving to represent the electric current being carried by a flow of the negative conduction electrons.

The point of the model in Figure (27-5) was to show that a left directed negative current, seen in (a) is essentially equivalent to a right directed positive current seen in (b). In Figure (27-5a), we drew a stick figure diagram of a person walking to the left at the same speed v as the negative rod. Figure (27-5b) is the same setup from the point of view of the stick figure person. She sees the negative rod at rest and the positive rod moving to the right as shown.



a) observer walking along with the moving negatively charged rod



 b) from the observer's point of view the negative rod is at rest and the positive charge is moving to the right

## Figure 27-5 a,b

In (a) we have a left directed negative current, while in (b) we have a right directed positive current. The only difference is the perspective of the observer. (You can turn a negative current into an oppositely flowing positive one simply by moving your head.) In another calculation, we saw that if a millimeter cross section copper wire carried a steady current of one ampere, the conduction electrons would have to move at the slow speed of 1/27 of a millimeter per second, a motion so slow that it would be hard to detect. As a result there should be no important physical difference between the two points of view, and a left directed negative current should be physically equivalent to a right directed positive current.

A closer examination of Figure (27-5) shows that we have left something out. The bottom figure, (27-5b) is not precisely what the moving observer sees. To show what has been left out, we have in Figure (2a) redrawn Figure (27-5a) and carefully labeled the individual charges. To maintain strict overall charge neutrality we have used charges +Q on the positive rod, charges -Q on the negative rod, and both sets of charges have equal separations of  $\ell$  centimeters.

From the point of view of the moving observer in Figure (2b), the negative rod is at rest and the positive rod is moving to the right as we saw back in Figure (27-5b). But, *due to the Lorentz contraction, the spacing between the charges is no longer*  $\ell$ ? Since the positive rod was at rest and is now moving, the length of the positive spacing must be contracted to a distance  $\ell \sqrt{1 - v^2/c^2}$  as shown.

On the other hand the negative rod was moving in Figure (2a), therefore the negative spacing must expand to  $\sqrt[4]{1 - v^2/c^2}$  when the negative rod comes to rest. (*Start with a spacing*  $\sqrt[4]{1 - v^2/c^2}$  for the negative charges at rest in Figure (2b), and go up to Figure (2a) where the negative rod is moving at a speed v. There the spacing must contract by a factor  $\sqrt{1 - v^2/c^2}$ , and the new spacing is  $\sqrt[4]{1 - v^2/c^2} \times \sqrt{1 - v^2/c^2} = \sqrt[6]{as shown.}$ 

As a result of the Lorentz contraction, the moving observer will see that the positive charges on her moving rod are closer together than the negative charges on her stationary rod. (We have exaggerated this effect in our sketch, Figure (2b)). Thus the moving observer of Figure (2b) sees not only a right directed positive current, but also *a net positive charge density* on her two rods. The Lorentz contraction has changed a neutral wire in Figure (2a) into a positively charged one in Figure (2b)!



a) Observer walking along with the moving negatively charged rod.



**b)** Charged rods from the observer's point of view. Now that the positive charge is moving, the spacing between positive charges has **contracted** from  $\ell$  to  $\ell \sqrt{1-v^2/c^2}$ . The negative rod is now at rest, the Lorentz contraction is undone, and the negative spacing has **expanded** from  $\ell$  to  $\ell / \sqrt{1-v^2/c^2}$ .

## Figure 2

An electric current from two points of view.

## **Charge Density on the Two Rods**

Our next step will be to calculate the net charge density  $\lambda$  on the pair of rods shown in Figure (2b). Somewhat messy algebra is required for this calculation, but the result will be used in much of the remainder of the text. The effort will be worth it.

If we have a rod with charges spaced a distance d apart as shown in Figure (3), then a unit length of the rod, 1 meter, contains 1/d charges. (For example, if d = .01 meter, then there will be 1/d = 100 charges per meter.) If each charge is of strength Q, then there is a total charge Q/d on each meter of the rod. Thus the charge density is  $\lambda = Q/d$  coulombs per meter. Applying this result to the positive rod of Figure (2b) gives us a positive charge density

$$\lambda_{+} = \frac{Q}{d_{+}} = \frac{Q}{\ell \sqrt{1 - v^{2}/c^{2}}}$$
(5)

And on the negative rod the charge density is

$$\lambda_{-} = \frac{-Q}{d_{-}} = \frac{-Q}{\ell/\sqrt{1 - v^{2}/c^{2}}}$$
$$= \frac{-Q\sqrt{1 - v^{2}/c^{2}}}{\ell}$$
(6)

Multiplying the top and bottom of the right side of Equation (6) by  $\sqrt{1 - v^2/c^2}$ , we can write  $\lambda_{-}$  as

$$\lambda_{-} = \frac{-Q\sqrt{1 - v^{2}/c^{2}}}{\ell} \times \frac{\sqrt{1 - v^{2}/c^{2}}}{\sqrt{1 - v^{2}/c^{2}}}$$
$$= \frac{-Q}{\ell\sqrt{1 - v^{2}/c^{2}}} \left(1 - v^{2}/c^{2}\right)$$
(7)

$$\frac{|\mathbf{k} - \mathbf{d} \rightarrow|}{\mathbf{Q} + \mathbf{Q} + \mathbf{Q} + \mathbf{Q}} = \frac{\lambda \text{ coulombs/meter}}{\lambda \text{ coulombs/meter}} = \frac{Q}{d}$$

#### Figure 3

If the charges are a distance d apart, then there are 1/d charges per meter of rod. (If d = .1 meters, then there are 10 charges/meter.) If the magnitude of each charge is Q, then  $\lambda$ , the charge per meter is Q times as great, i.e.,  $\lambda = Q * (1/d)$ .

The net charge density  $\lambda$  is obtained by adding  $\lambda_+$  and  $\lambda_-$  of Equations (5) and (7) to get

$$\lambda = \lambda_{+} + \lambda_{-} = \frac{Q}{\ell \sqrt{1 - v^{2}/c^{2}}} \left\{ 1 - (1 - v^{2}/c^{2}) \right\}$$
$$\lambda = \frac{Q}{\ell \sqrt{1 - v^{2}/c^{2}}} \left\{ \frac{v^{2}}{c^{2}} \right\} = \lambda_{+} \frac{v^{2}}{c^{2}}$$
(8)

Equation (8) can be simplified by noting that the current i carried by the positive rod in Figure (2b) is equal to the charge  $\lambda_+$  on 1 meter of the rod times the speed v of the rod

$$i = \lambda_{+}v$$
 carried by the (9)  
positive rod

(In one second, v meters of rod move past any fixed cross-sectional area, and the charge on this v meters of rod is  $\lambda_{+}$  v) Using Equation (9), we can replace  $\lambda_{+}$  and one of the v's in Equation (8) by i to get the result

$$\lambda = \frac{iv}{c^2}$$
(10)

Due to the Lorentz contraction, the moving observer in Figure (2b) sees a net *positive charge density*  $\lambda = iv/c^2$  on the wire which from our point of view, Figure (2a) *was precisely neutral*.

Although Equation (10) may be formally correct, one has the feeling that it is insane to worry about the Lorentz contraction for speeds as slow as 2 millimeters per minute. But the Lorentz contraction changes a precisely neutral pair of rods shown in Figure (2a), into a pair with a net positive charge density  $\lambda = iv/c^2$  in Figure (2b). We have unbalanced a perfect cancellation of charge which could lead to an imbalance in the cancellation of electrostatic forces. Since we saw from our discussion of the two garden peas that imbalances as small as one part in  $10^{18}$  or less might be observable, let us see if there are any real experiments where the charge density  $\lambda$  is detectable.

## **A Proposed Experiment**

How would we detect the charge imbalance in Figure (2b)? If there is a net positive charge density l on the two rods in Figure (2b), repeated here again in Figure (4), then the net charge should produce a radial electric field whose strength is given by the formula

$$E = \frac{\lambda}{2\pi\varepsilon_0 r}$$
(11)

We derived this result in our very first discussions of Coulomb's law in Chapter 24. (*Remember that the two* separate rods are our model for a single copper wire carrying a current. The rods are not physically separated as we have had to draw them, the negative conduction electrons and positive nuclei are flowing through each other.)

We can test for the existence of the electric field produced by the positive charge density  $\lambda = iv/c^2$  by placing a test particle of charge q a distance r from the wire as shown in Figure (4). This test particle should experience a force

$$\vec{F} = q\vec{E} \tag{12}$$

which would be repulsive if the test particle q is positive and attractive if q is negative. Using Equations (10) for  $\lambda$  and (11) for E, Equation (12) gives for the predicted magnitude of  $\vec{F}$ :

$$\left|\vec{F}\right| = q\left|\vec{E}\right| = \frac{q\lambda}{2\pi\epsilon_0 r} = \frac{q\,iv}{2\pi\epsilon_0 rc^2}$$
 (13)

Rearranging the terms on the right side of Equation (13), we can write  $|\vec{F}|$  in the form

$$\left|\vec{F}\right| = qv \times \left\{\frac{i}{2\pi r\epsilon_0 c^2}\right\}$$
 (14)

Why we have written Equation (14) this way will become clear shortly.



#### Figure 4

To test for the net charge density, as seen by the observer at rest relative to the minus charge, the observer places a test charge q a distance r from the wire as shown. If there is a net charge  $\lambda$  on the wire, the charge will produce an electric field  $\vec{E}$ , which will exert a force  $\vec{F} = q\vec{E}$  on the test particle as shown.

# **Origin of Magnetic Forces**

You might think that the next step is to put reasonable numbers into Equation (14) and see if we get a force  $\vec{F}$  that is strong enough to be observed. But there is an important thought experiment we will carry out first. The idea is to look at the force on a test particle from two different points of view, one where the wire appears charged as in Figures (4 & 2b), and where the wire appears neutral as in Figure (2a). The two points of view are shown in Figure (5).

Figure (5b), on the left, is the situation as observed by the moving observer. She has a copper wire carrying a positive current directed to the right. Due to the Lorentz contraction, her copper wire has a charge density  $\lambda$  which creates an electric field  $\vec{E}$ . To observe  $\vec{E}$ , she mounts a test particle -q at one end of a spring whose other end is fixed, nailed to her floor. She detects the force  $\vec{F} = -q\vec{E}$  by observing how much the spring has been stretched.

*Our* point of view is shown in Figure (5a). It is exactly the same setup, we have touched nothing! It is just viewed by someone moving to the right relative to her.

In our point of view, the moving observer, the negative rod, and the test particle are all moving to the left at a

speed v. The positive rod is at rest, the Lorentz contractions are undone, and there is *no net charge* on our rods. All we have is a negative current flowing to the left.

We can also see the test particle. It is now moving to the left at a speed v, and it is still attached to the spring.

Here is the crucial point of this discussion. We also see that the spring is stretched. We also see that the end of the spring has been pulled beyond the mark indicating the unstretched length. We also detect the force  $\vec{F}$  on the test particle!

Why do we see a force  $\vec{F}$  on the test particle? Our copper wire is electrically neutral; we do not have an electric field  $\vec{E}$  to produce the force  $\vec{F}$ . Yet  $\vec{F}$  is there. If we cut the spring, the test particle would accelerate toward the copper wire, and both we and the moving observer would see this acceleration.

At this point, we have come upon a basic problem. Even if the Lorentz contraction is very small and the force  $\vec{F}$  in Figure (5b) is very small, we at least predict that  $\vec{F}$  exists. In Figure (5a) we predict that a neutral wire, that is carrying a current but *has absolutely no net charge on it*, exerts an attractive force on a moving negative charge as shown.



#### Figure 5

Two views of the same experiment. For the observer moving with the electrons, she sees a positively charged wire exerting an attractive force on the negative charge at rest. We see an electrically neutral wire carrying a negative current, and a moving negative charge. The spring is still stretched, meaning the attractive force is still there.

With a few modifications, the experiment shown in Figure (5a) is easy to perform and gives clear results. Instead of a negative test particle attached to a spring, we will use a beam of electrons in an electron gun as shown in Figure (6). In Figure (6a) we see the setup of our thought experiment. In Figure (6b) we have replaced the two charged rods with a neutral copper wire carrying a current -i, and replaced the test particle with an electron beam.

According to Equation (14), the force  $\vec{F}$  on the test particle -q should have a strength proportional to the current i in the wire. Thus when we turn on a current (shorting the wire on the terminals of a car storage

*battery to produce a healthy current)* we will see the electron beam deflected toward the wire if there is an observable force. The experimental result is shown in Figure (6c). There is a large, easily observed deflection. The force  $\vec{F}$  is easily seen.

#### **Exercise 2**

In Figure (7) we reversed the direction of the current in the wire and observe that the electron beam is deflected away from the wire. Devise a thought experiment, analogous to the one shown in Figure (5a,b) that explains why the electron beam is repelled from the wire by this setup. (This is not a trivial problem; you may have to try several charge distributions on moving rods before you can imitate the situation shown in Figure (7a). But the effort is worth it because you will be making a physical prediction that is checked by the experimental results of Figure (7b).



#### Figure 6c

For an experimental test of the results of the thought experiment, we replace the moving negative charge with a beam of electrons in an electron gun. The electrons are attracted to the wire as predicted.

# Figure 7

If we reverse the direction of the current in the wire, the electrons in the beam are repelled

## **Magnetic Forces**

Historically an electric force was defined as the force between charged particles and was expressed by Coulomb's law. The force in Figure (5a) between a moving test charge and an *uncharged* wire does not meet this criterion. You might say that for historical reasons, it is not eligible to be called an electric force.

The forces we saw in Figures (6c) and (7b), between a moving charge and a neutral electric current, were known before special relativity and were called *magnetic forces*. Our derivation of the magnetic force in Figure (5a) from the electric force seen in Figure (5b) demonstrates that *electric and magnetic forces in this example are the same thing just seen from a different point of view*.

When we go from Figure (5b) to (5a), which we can do by moving our head at a speed of 2 millimeters per minute, we see essentially no change in the physical setup but we have an enormous change in perspective. We go from a right directed positive current to a left directed negative current, and the force on the test particle changes from an electric to a magnetic force.



#### Figure 8

Force on a charge -q moving at a speed v parallel to a negative current -i a distance r away.

# **MAGNETIC FORCE LAW**

From our Coulomb's law calculation of the electric force in Figure (5a), we were able to obtain the formula for the magnetic force in Figure (5b). The result, Equation (14) repeated here, is

$$\left|\vec{\mathbf{F}}\right| = q\mathbf{v} \times \left\{\frac{i}{2\pi r\epsilon_0 c^2}\right\}$$
 (14)

where q is the charge on the test particle, i the current in the wire, and r the distance from the wire to the charge as shown in Figure (8). The only thing our derivation does not make clear is whether v in Equation (14) is the speed of the test charge or the speed of the electrons in the wire. We can't tell because we used the same speed v for both in our thought experiment. A more complex thought experiment will show that the v in Equation (14) is the speed of the test particle.

# The Magnetic Field B

In Equation (14) we have broken the somewhat complex formula for the magnetic force into two parts. The first part qv is related to the test charge (q is its charge and v its speed), and the second part in the curly brackets, which we will designate by the letter B

$$B \equiv \frac{i}{2\pi r \varepsilon_0 c^2} \tag{15}$$

is related to the wire. The wire is carrying a current i and located a distance r away.

The quantity B in Equation (15) is called the magnitude of *the magnetic field of the wire*, and in terms of B the magnetic force becomes

$$\left|\vec{F}_{magnetic}\right| = qvB$$
 (16)

Equation (16) is almost a complete statement of the magnetic force law. What we have left to do for the law is to assign a direction to B, i.e. turn it into the vector  $\vec{B}$ , and then turn Equation (16) into a vector equation for the force  $\vec{F}_{magnetic}$ .

There is one more definition. In the MKS system of units, it is traditional to define the constant  $\mu_0$  by the equation

$$\mu_0 = \frac{1}{\varepsilon_0 c^2} \qquad definition of \mu_0 \tag{17}$$

Using this definition of  $\mu_0$  in Equation (15) for B, we get

$$B = \frac{\mu_0 i}{2\pi r} \qquad magnetic field of a wire \qquad (18)$$

as the formula for the magnetic field of a wire.

It turns out to be quite an accomplishment to get Equations (16), (17), and (18) out of one thought experiment. These equations will provide the foundation for most of the rest of our discussion of electric and magnetic (electromagnetic) theory.

# **Direction of the Magnetic Field**

We will temporarily leave our special relativity thought experiment and approach magnetism in a more traditional way. Figure (9) is a sketch of the magnetic field of the earth. By convention the direction of the magnetic field lines are defined by the direction that a compass needle points. At the equator the magnetic field lines point north (as does a compass needle) and the field lines are parallel to the surface of the earth. As we go north from the equator the magnetic field lines begin to point down into the earth as well as north. At the north magnetic pole the magnetic field lines go straight down.

Figure (9) is drawn with the magnetic north pole at the top. The earth's rotational axis, passing through the true north pole, is at an angle of 11.5 degrees as shown. Over time the location of the earth's magnetic pole wanders, and occasionally flips down to the southern hemisphere. Currently the north magnetic pole is located in north central Canada.



#### Figure 9

Magnetic field of the earth. The magnetic field lines show that the direction a freely floating compass needle would point at any location outside the earth. For example, at the equator the compass needle would be parallel to the surface of the earth and point north. At the north magnetic pole, the compass needle would point straight down (and thus not be very useful for navigation).

As we mentioned, it is by long standing convention that the direction of the magnetic field is defined by the direction a compass needle points. We can therefore use a set of small compasses to map the direction of the magnetic field.

In 1820, while preparing a physics lecture demonstration for a class of students, Hans Christian Oersted discovered that an electric current in a wire could



**Figure 10a** With no current flowing in the wire, all the compass needles point north.



**Figure 10b** When an upward directed current is turned on, the compass needles point in a counterclockwise circle about the wire.

deflect a compass needle. This was the first evidence of the connection between the subject of electricity with its charges and currents, and magnetism with its magnets and compasses.

The fact that a wire carrying a current deflects a compass needle means that the current must be producing a magnetic field. We can *use the deflected compass needles to show us the shape of the magnetic field* of a wire. This is done in Figures (10a,b) where we see a ring of compasses surrounding a vertical wire. In (10a) there is no current in the wire, and all the compass needles all point north (black tips). In (10b) we have turned on an upward directed current in the wire, and the compass needles point in a circle around the wire. *Using the north pole of the compass needle to define the direction of the magnetic field*, we see that the magnetic field goes in a counterclockwise circle around the wire.

In Figure (11) we have replaced the compasses in Figure (10) with a sprinkle of iron filings. When the current in the wire is turned on, the iron filings align themselves to produce the circular field pattern shown. What is happening is that each iron filing is acting as a small compass needle and is lining up parallel to the magnetic field. While we cannot tell which way is north with iron filings, we get a much more complete



**Figure 11** Iron fillings sprinkled around a current form a circular pattern. Each iron filing lines up like a compass needle, giving us a map of the magnetic field.

picture or map of the direction of the magnetic field. Figure (11) is convincing evidence that the magnetic field surrounding a wire carrying a current is in a circular field, not unlike the circular flow pattern of water around the core of a vortex.

The use of iron filings turns out to be a wonderfully simple way to map magnetic field patterns. In Figure (12), a sheet of cardboard was placed on a bar magnet and iron filings sprinkled on the cardboard. The result, with two poles or points of focus resembles what is called a *dipole field*.

In Figure (13) we have thrown iron filings at an old iron magnet and created what one young observer called a "magnet plant." Here we see the three dimensional structure of the magnetic field, not only between the pole pieces but over the top half of the magnet.



#### Figure 12

A sheet of cardboard is placed over the poles of a magnet and sprinkled with iron filings. From the pattern of the filings we see the shape of the more complex magnetic field of the magnet.



#### Figure 13

You get a three dimensional picture of the magnetic field if you pour the iron filings directly on the magnet. Our young daughter called this a Magnet Plant.

# The Right Hand Rule for Currents

Iron filings give us an excellent picture of the shape of the magnetic field, but do not tell us which way the field is pointing. For that we have to go back to compasses as in Figure (9), where  $\vec{B}$  is defined as pointing in the direction of the north tip of the compass needle. In that figure we see that when a positive current i is flowing toward us, the magnetic field goes in a counter clockwise direction as illustrated in Figure (14).

The above description for the direction may be hard to remember. A more concise description is the following. **Point the thumb of your right hand in the direction of the current** as shown in Figure (14), then **your fingers will curl in the direction of the magnetic field**. This mnemonic device for remembering the direction of  $\vec{B}$  is one of the **right hand rules**. (This is the version we used in Figure 2-37 to distinguish right and left hand threads.) If we had used compasses that pointed south, we would have gotten a left hand rule.



#### Figure 14

Right hand rule for the magnetic field of a current i. Point the thumb in the direction of the positive current and your fingers curl in the direction of the magnetic field.

## **Parallel Currents Attract**

While we are in the business of discussing mnemonic rules, there is another that makes it easy to remember whether a charge moving parallel to a current is attracted or repelled. In Figure (6) we had a beam of negative electrons moving parallel to a negative current -i, and the electrons were attracted to the current. In Figure (7) the current was reversed and the electrons were repelled. One can work out a thought experiment similar to the ones we have done in this chapter to show that a positive charge moving parallel to a positive current as shown in Figure (15) is attracted.

The simple, yet general rule is that **parallel currents attract, opposite currents repel**. A positive charge moving in the direction of a positive current, or a negative charge moving along with a negative current are attracting parallel currents. When we have negative charges moving opposite to a negative current as in Figure (7) we have an example of opposite currents that repel.



#### Figure 15

A positive charge, moving parallel to a positive current, is attracted by the current. Thinking of the moving positive charge as a positive upward directed current, we have the rule that parallel currents attract, opposite currents repel.

## The Magnetic Force Law

Now that we have a direction assigned to the magnetic field  $\vec{B}$  we are in a position to include directions in our formula for magnetic forces. In Figure (16) which is the same as (15) but also shows the magnetic field, we have a positive charge moving parallel to a positive current, and therefore an attractive force whose magnitude is given by Equation (16) as

$$\left|\vec{F}_{mag}\right| = qvB \tag{16}$$

There are three different vectors in Equation (16),  $\vec{F}_{mag}$ ,  $\vec{v}$ , and  $\vec{B}$ . Our problem is to see if we can combine these vectors in any way so that something like Equation (16) tells us both the magnitude and the direction of the magnetic force  $\vec{F}_{mag}$ . That is, can we turn Equation (16) into a vector equation?





The right hand side of Equation (16) involves the product of the vectors  $\vec{v}$  and  $\vec{B}$ . So far in the text we have discussed two different ways of multiplying vectors; the dot product  $\vec{A} \cdot \vec{B}$  which gives a scalar number C, and the cross product  $\vec{A} \times \vec{B}$  which gives the vector  $\vec{C}$ . Since we want the product of  $\vec{v}$  and  $\vec{B}$  to give us the vector  $\vec{F}_{mag}$ , the cross product appears to be the better candidate, and we can try

$$\vec{F}$$
mag =  $\vec{qv} \times \vec{B}$  magnetic  
force law (19)

as our vector equation.

To see if Equation (19) works, look at the three vectors  $\vec{v}$ ,  $\vec{B}$ , and  $\vec{F}_{mag}$  of Figure (16) redrawn in Figure (17). The force  $\vec{F}_{mag}$  is perpendicular to the plane defined by  $\vec{v}$  and  $\vec{B}$  which is the essential feature of a vector cross product. To see if  $\vec{F}_{mag}$  is in the correct direction, we use the cross product right hand rule. Point the fingers of your *right* hand in the direction of the first vector in the cross product, in this case  $\vec{v}$ , and curl them in the direction of the second vector, now  $\vec{B}$ . Then your thumb will point in the direction of the cross product  $\vec{v} \times \vec{B}$ . Looking at Figure (17), we see that the thumb of the right hand sketch does point in the direction of  $\vec{F}_{mag}$ , therefore the direction of  $\vec{F}_{mag}$  is correctly given by the cross product  $\vec{v} \times \vec{B}$ . (If the direction had come out wrong, we could have used  $\vec{B} \times \vec{v}$  instead.)



#### Figure 17

Right hand rule for the vector cross product  $\vec{v} \times \vec{B}$ . Point the fingers of your right hand in the direction of the first vector  $\vec{v}$ , and then curl them in the direction of the second vector  $\vec{B}$ . Your thumb ends up pointing in the direction of the vector  $\vec{v} \times \vec{B}$ .

Although the formula for  $\vec{F}_{mag}$ , Equation (19), was derived for a special case, the result is general. Whenever a particle of charge q is moving with a velocity  $\vec{v}$  through a magnetic field  $\vec{B}$ , no matter what the relative directions of  $\vec{v}$  and  $\vec{B}$ , the magnetic force is correctly given as  $q\vec{v} \times \vec{B}$ .

#### **Exercise 3**

Using the magnetic force law  $\vec{F}_{mag} = q\vec{v} \times \vec{B}$  and the right hand rule for the magnetic field of a current, show that:

(a) An electron moving parallel to a negative current -i is attracted (Figure 6)

(b) An electron moving opposite to a negative current is repelled (Figure 7)

#### **Lorentz Force Law**

Since electric and magnetic forces are closely related, it makes sense to write one formula for both the electric and the magnetic force on a charged particle. If we have a charge q moving with a velocity  $\vec{v}$  through an electric field  $\vec{E}$  and a magnetic field  $\vec{B}$ , then the electric force is  $q\vec{E}$ , the magnetic force  $q\vec{v} \times \vec{B}$ , and the total "electromagnetic" force is given by

$$\vec{F} = q\vec{E} + q\vec{v} \times \vec{B}$$
*Lorentz force law*
(20)

Equation (20), which is known as the *Lorentz force law*, is a complete description of the electric and magnetic forces on a charged particle, provided  $\vec{E}$  and  $\vec{B}$  are known.

# Dimensions of the Magnetic Field, Tesla and Gauss

The dimensions of the magnetic field can be obtained from the magnetic force law. In the MKS system we have

$$F(newtons) = q(coulombs)v\left(\frac{meters}{second}\right) \times B$$

which gives us B in units of newton seconds per coulomb meter. This set of dimensions is given the name *tesla* 

$$\frac{\text{newton second}}{\text{coulomb meter}} \equiv \text{tesla} \begin{cases} MKSunits \\ for \\ magnetic \\ fields \end{cases}$$
(21)

Although most MKS electrical quantities like the volt and ampere are convenient, the tesla is too large. Only the strongest electromagnets, or the new superconducting magnets used in particle accelerators or magnetic resonance imaging apparatus, can produce fields of the order of 1 tesla or more. Fields produced by coils of wire we use in the lab are typically 100 times weaker, and the earth's magnetic field is 100 times weaker still.

In the CGS system of units, magnetic fields are measured in gauss, where

$$1 \text{ gauss} = 10^{-4} \text{ tesla}$$

The gauss is so much more convenient a unit that there is a major incentive to work with CGS units when studying magnetic phenomena. For example the earth's magnetic field has a strength of about 1 gauss at the earth's surface, and the magnetic field that deflected the electrons in Figures (6) and (7) has a strength of about 30 gauss at the electron beam. Refrigerator magnets have comparable strengths.

We could be pedantic, insist on using only MKS units, and suffer with numbers like .00021 tesla in discussions of the earth's magnetic field. But if someone wants you to measure a magnetic field, they hand you a "gauss meter" not a tesla meter. Magnetic-type instruments are usually calibrated in gauss. If you worked only with tesla, you would have a hard time communicating with much of the scientific community. What we will do in this text is use either gauss or tesla depending upon which is the more convenient unit. When we come to a calculation, we will convert any gauss to tesla, just as we convert any distances measured in centimeters to meters.

# **Uniform Magnetic Fields**

Using the magnetic force law  $\vec{F}_{mag} = q\vec{v} \times \vec{B}$  to calculate magnetic forces is often the easy part of the problem. The hard part can be to determine the magnetic field  $\vec{B}$ . For a current in a straight wire, we were able to use a thought experiment and the Lorentz contraction to get Equation (20) for the strength of  $\vec{B}$ . But in more complicated situations, where we may have bent wires, thought experiments become too difficult and we need other techniques for calculating  $\vec{B}$ .

One of the other techniques, which we will discuss in the next chapter, is called *Ampere's law*. This law will give us the ability to calculate the magnetic field of simple current distribution much the same way that Gauss' law allowed us to calculate the electric field of simple charge distributions. But until we get to Ampere's law in the next chapter, we will confine our study of the magnetic force law to the simplest of all possible magnetic fields, the uniform magnetic field.



**Figure 18** Between the poles of this magnet there is a relatively uniform magnetic field.

Working with uniform magnetic fields, fields that are constant in both magnitude and direction, is so convenient that physicists and engineers go to great lengths to construct them. One place to find a uniform field is between the flat pole pieces of a magnet, as seen in Figure (18) which is our "magnet plant" of Figure (13) with fewer iron filings.

If we bend a wire in a loop, then a current around the loop produces the fairly complex field pattern shown in Figure (19). When we use two loops as seen in Figure (20), the field becomes more complicated in some places but begins to be more uniform in the central region between the coils. With many loops, with the coil of wire shown in Figure (21a), we get a nearly uniform field inside. Such a coil is called a solenoid, and will be studied extensively in the next chapter. An iron filing map of the field of a large diameter, tightly wound solenoid is seen in Figure (21b).



**Figure 19** *The magnetic field of a current loop is fairly complex.* 



#### Figure 21a

Magnetic field in the upper half of a section of a coil of wire. When you have many closely spaced coils, the field inside can become quite uniform through most of the length of the coil.



#### Figure 20

The magnetic field in the region between a pair of coils is relatively uniform. We can achieve the greatest uniformity by making the separation d between the coils equal to the radius of the coils. Such a setup is called a pair of Helmholtz coils.



**Figure 21b** *Iron filing map of the magnetic field of a large diameter coil. (Student project, Alexandra Lesk and Kirsten Teany.)* 

# **Helmholtz Coils**

For now we will confine our attention to the reasonably uniform field in the central region between two coils seen in Figure (20). Helmholtz discovered that when the coils are spaced a distance d apart equal to the coil radius r (Figure 22), we get a maximally uniform field B between the coils. This arrangement, which is called a pair of *Helmholtz coils*, is commonly used in physics and engineering apparatus. Figure (23) shows a pair of Helmholtz coils we use in our undergraduate physics labs and which will be used for several of the experiments discussed later. An iron filing map of the field produced by these coils is seen in Figure (24a), and one of the experiments will give us a field plot similar to Figure (24b).

In our derivation of the magnetic field of a current in a straight wire, we saw that *the strength of the magnetic field was proportional to the current i in the wire*. This is true even if the wire is bent to form coils, or even twisted into a complex tangle. That means that once you have mapped the magnetic field for a given current (i) in a set of wires, doubling the current produces the same shape map with twice as strong a field.



**Figure 22** For Helmholtz coils, the separation d equals the coil radius r.

For the Helmholtz coils in Figure (23), it was observed that when a current of one amp flowed through the coils, the strength of the magnetic field in the central regions was 8 gauss. A current of 2 amps produced a 16 gauss field. Thus the field strength, **for these coils**, is related to the current i by

$$B(gauss) = 8i(amps)$$
 for the Helmholtz  
coils of  
Figure (23) only

In the lab we measure the strength of B simply by reading (i) from an ampmeter and multiplying by 8. Of course, if you are using a different set of coils,(i) will be multiplied by a different number. (*Do not worry about the mixed units, remember that we convert gauss to tesla before doing MKS calculations.*)

One can derive a formula for an idealized set of Helmholtz coils. The derivation is complicated and the answer B =  $8\mu_0 \text{Ni}/(5\sqrt{5} \text{ r})$  is rather a mess. (N is the number of turns in each coil.) The simple feature which we expected, is that B is proportional to the strength of the current (i) in the coils. Because another law, called Faraday's law, can be used to give us a more accurate calibration of real Helmholtz coils, we will leave the derivation of the Helmholtz formula above to other texts.



**Figure 23** Helmholtz coils used in a number of lab experiments discussed in the text. Each coil consists of 60 turns of fairly heavy magnet wire.



#### Figure 24a

Iron filing map of the magnetic field of the Helmholtz coils. (Student project, Alexandra Lesk and Kirsten Teany.)



#### Figure 24b

Plot from a student experiment, of the magnetic field in the region between and around the coils.

# MOTION OF CHARGED PARTICLES IN MAGNETIC FIELDS

In physics, one of the primary uses of magnetic fields is to control the motion of charged particles. When the magnetic field is uniform, the motion is particularly simple and has many practical applications from particle accelerators to mass spectrometers. Here we will discuss this motion and several of the applications.

The main feature of the magnetic force law,

$$\mathbf{F}_{\text{magnetic}} = \mathbf{q} \vec{\mathbf{v}} \times \mathbf{B} \tag{19}$$

is that because of the cross product  $\vec{v} \times \vec{B}$ , the magnetic force is *always* perpendicular to the velocity  $\vec{v}$  of the charged particle. This has one important immediate consequence. *Magnetic forces do no work!* The formula for the power, i.e., the work done per second, is

Work done  
per second  
by a force 
$$\vec{F}$$
 = power =  $\vec{F} \cdot \vec{v}$  (22)

Since the magnetic force  $\vec{F}_{mag}$  is always perpendicular to  $\vec{v}$ , we have

$$\begin{array}{l} \text{Work done by a} \\ \text{magnetic force} \end{array} \end{array} = \vec{F}_{\text{magnetic}} \cdot \vec{v} \\ = q(\vec{v} \times \vec{B}) \cdot \vec{v} \equiv 0 \end{array}$$
(23)

magnetic fields do not change the energy of a particle, they simply change the direction of motion.



## Figure 25a

Top view looking down on the electron gun placed between the Helmholtz coils. The electrons in the beam move perpendicular to the magnetic field  $\vec{B}$ . The magnetic force  $\vec{F} = (-e) \vec{v} \times \vec{B}$  is directed up, out of the paper in this drawing.



#### Figure 25b

Side view of the electron beam, as seen through the lower coil in Figure 25a. In this view the magnetic field is directed out of the paper toward the reader.



**Figure 25d** *Movie of the experiment.* 

# Motion in a Uniform Magnetic Field

When we have a charged particle moving through a uniform magnetic field, we get a particularly simple kind of motion—the circular motion seen in Figure (25b). In Figure (25a) we sketched the experimental setup where an electron gun is placed between a pair of Helmholtz coils so that the magnetic field  $\vec{B}$  is perpendicular to the electron beam as shown. Figure (25b) is a photograph of the electron beam deflected into a circular path. In Figure (25c) we have a sketch of the forces on an electron in the beam. The magnetic field B in this diagram is up out of the paper, thus  $\vec{v} \times \vec{B}$  points radially out from the circle. But the electron has a negative charge, thus the magnetic force  $\vec{F}_{B}$ 

$$\vec{\mathbf{F}}_{\mathbf{B}} = (-\mathbf{e})\vec{\mathbf{v}}\times\vec{\mathbf{B}}$$

points in toward the center of the circle as shown.



## Figure 25c

As the electrons move along a curved path, the magnetic force  $\vec{F} = -q \vec{v} \times \vec{B}$  always remains perpendicular to the velocity and therefore cannot change the speed of the electrons. The resulting motion is uniform circular motion where the force and the acceleration are directed toward the center of the circle. To apply Newton's second law to the electrons in Figure (25), we note that a particle moving in a circle accelerates toward the center of the circle, the same direction as  $\vec{F}_{mag}$  in Figure (25c). Thus  $\vec{F}_B$  and  $m\vec{a}$  are in the same direction and we can use the fact that for circular motion  $a = v^2/r$  to get  $|\vec{F}_B| = m|\vec{a}|$  or

$$qvB = mv^2/r \tag{24}$$

Solving for r, we predict from Equation (24) that the electron beam will be bent into a circle of radius r given by

$$r = \frac{mv}{qB}$$
(25)

Equation (25) is an important result that we will use often. But it is so easy to derive, and it is such good practice to derive it, that it may be a good idea not to memorize it.

Let us use the experimental numbers provided with Figure (25b) as an example of the use of Equation (25). In that figure, the strength of the magnetic field is B = 70 gauss, and the electrons were accelerated by an accelerating voltage of 135 volts. The constants m and q are the mass and charge of an electron.

The first step is to calculate the speed v of the electrons using the fact that the electrons have 135 eV of kinetic energy. We begin by converting from eV to joules using the conversion factor  $1.6 \times 10^{-19}$  joules per eV. This gives

$$\frac{1}{2}mv^{2} = 135 \text{ eV} \times 1.6 \times 10^{-19} \frac{\text{joules}}{\text{eV}}$$
$$v^{2} = \frac{2 \times 135 \times 1.6 \times 10^{-19}}{.911 \times 10^{-30}}$$
$$= 47.4 \times 10^{12} \frac{\text{m}^{2}}{\text{s}^{2}}$$
$$v = 6.9 \times 10^{6} \frac{\text{m}}{\text{s}}$$

Next convert B from gauss to the MKS tesla

$$70 \text{ gauss} = 70 \times 10^{-4} \text{ tesla}$$
 (27)

Substituting Equations (26) and (27) in (25) gives

$$r = \frac{mv}{qB} = \frac{.911 \times 10^{-30} \times 6.9 \times 10^{6}}{1.6 \times 10^{-19} \times 7 \times 10^{-3}}$$
$$r = 5.6 \times 10^{-3} \text{ m} = .56 \text{ cm}$$
(28)

#### Exercise 4

The scale of distance shown in Figure (25b) was drawn knowing the dimensions of the cap in the electron gun. Use this scale to estimate the radius of curvature of the electron beam and compare the result with the prediction of Equation (28).

#### **Exercise 5**

Use the experimental results shown in Figure (26) (B = 70 gauss) to estimate the accelerating voltage used for the electrons in the beam. (The experimental answer is included in the homework answer section.)



#### Figure 26

Use the fact that the magnetic field for this example was 70 gauss to estimate the accelerating voltage that produced the electron beam.

# **Particle Accelerators**

Our knowledge of the structure of matter on a subatomic scale, where we study the various kinds of elementary particles, has come from our ability to accelerate particles to high energies in particle accelerators such as the synchrotron. In a synchrotron, an electric field  $\vec{E}$  is used to give the particle's energy, and a magnetic field  $\vec{B}$  is used to keep the particles confined to a circular track.

Figure (27a) is a schematic diagram of a small electron synchrotron. At the top is an electron gun that is used to produce a beam of electrons. In practice the gun is quickly turned on then off to produce a pulse of electrons.



cross-sectional view

## Figure 27a

Diagram of a synchrotron, in which the electrons, produced by the electron gun, travel through a circular evacuated doughnut. The electrons are accelerated by an electric field, gaining energy on each trip around. The electrons are kept in a circular orbit by an increasingly strong magnetic field produced by the electromagnets. The pulse of electrons enter an evacuated circular track shown in the top view. To keep the pulse of electrons moving in the circular track, large electromagnets shown in the cross-sectional view are used to provide a perpendicular magnetic field. In this example the magnetic field  $\vec{B}$  points downward so that the magnetic force  $q\vec{v} \times \vec{B} = -e\vec{v} \times \vec{B}$  points inward toward the center of the track.

We saw that a magnetic field cannot do any work on the electrons since the magnetic force is perpendicular to the particle's velocity. Therefore to give the electrons more energy, we use an electric field  $\vec{E}$ . This is done by inserting into one section of the path a device that produces an electric field so that the electric force  $-e\vec{E}$  points in the direction of the motion of the electrons.

One might think of using a charged parallel plate capacitor to create the electric field  $\vec{E}$ , but that is not feasible. Later we will see that radio waves have an electric field  $\vec{E}$  associated with them, and it is a radio wave electric field in a so-called "resonant cavity" that is used to produce the required strong fields. For now it does not matter how  $\vec{E}$  is produced, it is this electric field that adds energy to the electrons.



## Figure 27b

The Berkeley synchrotron shown here, accelerated protons rather than electrons. It was the first machine with enough energy to create anti protons. After this machine was built, ways were devised for focusing the particle beam and using an evacuated doughnut with a much smaller cross-sectional area. When electrons gain energy, their momentum p = mv increases. Writing Equation (25) in the form

$$r = \frac{mv}{qB} = \frac{p}{qB}$$
(25a)

we see that an increase in the electron's momentum p will cause the orbital radius r to increase. The radius r will increase unless we compensate by increasing the strength B of the magnetic field. The rate at which we increase B must be synchronized with the rate at which we increase the particle's momentum p in order to keep r constant and keep the electrons in the circular path. Because of this synchronization, the device is called a *synchrotron*.

You can see that the amount of energy or momentum we can supply to the particles is limited by how strong a field B we can make. Iron electromagnets can create fields up to about 1 tesla (here the MKS unit is useful) or 10,000 gauss. The superconducting magnets, being used in the latest accelerator designs, can go up to around 5 tesla.

Noting that B is limited to one or a few tesla, Equation (25) tells us that to get more momentum or energy, we must use accelerators with a bigger radius r. This explains why particle accelerators are getting bigger and bigger. The biggest particle accelerator now operating in the United States is the proton accelerator at the Fermi National Accelerator Laboratory in Batavia, Illinois shown in Figures (28) and (29).

In Figure (28), we see a section of tunnel and the magnets that surround the 2 inch diameter evacuated pipe which carries the protons. Originally there was one ring using iron magnets (painted red and blue in the photograph). Later another ring with superconducting magnets was installed, in order to obtain stronger magnetic fields and higher proton energies. The ring of superconducting magnets (painted yellow) is beneath the ring of iron magnets.

Figure (29) is an aerial view showing the 4 mile circumference of the accelerator. Currently the largest accelerator in the world is at the European Center for Particle Physics (CERN). The 27 kilometer path of that accelerator is seen in Figure (30) on the next page.



#### Figure 28

The Fermi Lab accelerator has two accelerating rings, one on top of the other. In each, the evacuated doughnut is only 2 inches in diameter, and four miles in circumference. The bottom ring uses superconducting magnets (painted yellow), while the older upper ring has iron magnets (painted red and blue.)



**Figure 29** *Aerial view of the Fermi Lab particle accelerator.* 

# RELATIVISTIC ENERGY AND MOMENTA

Even the smallest synchrotrons accelerate electrons and protons up to relativistic energies where we can no longer use the non relativistic formula  $1/2 \text{ mv}^2$  for kinetic energy. For any calculations involving the large accelerators we must use fully relativistic calculations like  $E = \text{mc}^2$  for energy and p = mv for momentum where  $m = m_0/\sqrt{1 - v^2/c^2}$  is the relativistic mass.

Equation (25) or (25a) for a charged particle moving in a circular orbit of radius r, can be written in the form

$$\mathbf{p} = \mathbf{q}\mathbf{B}\mathbf{r} \tag{25b}$$

where B is the strength of the uniform magnetic field and p the particle momentum. It turns out that Equation (25) is correct even at relativistic energies provided p = mv is the relativistic momentum. Thus a knowledge of the magnetic field and orbital radius immediately tells us the momentum of the particles in the large synchrotrons.

To determine the energy of the particles in these machines, we need a relationship between a particle's energy E and momentum p. The relationship can be obtained by writing out E and p in the forms

$$p = mv = \frac{m_0}{\sqrt{1 - v^2/c^2}} v$$
 (29)

$$E = mc^{2} = \frac{m_{0}}{\sqrt{1 - v^{2}/c^{2}}} c^{2}$$
(30)

It is then straightforward algebraic substitution to show that

$$E^{2} = p^{2}c^{2} + m_{0}^{2}c^{4}$$
 An exact  
relationship (31)



#### Figure 30

Path for the 8 kilometer circumference Super Proton Synchrotron (SPS, solid circle) and the 27 kilometer Large Electron-Positron collider (LEP, dashed circle) at CERN, on the border between France and Switzerland. The Geneva airport is in the foreground.

#### **Exercise 6**

Directly check Equation (31) by plugging in the values of p and E from Equations (29) and (30).

In the big particle accelerators the kinetic energy supplied by the accelerators greatly exceeds the particle's rest energy  $m_0 c^2$ , so that the  $(m_0 c^2)^2$  term in Equation (31) is completely negligible. For these "highly relativistic" particles, we can drop the  $m_0^2 c^4$  term in Equation (31) and we get the much simpler formula

$$E \approx pc$$
 If  $E \gg m_0 c^2$  (32)

Equation (32) is an accurate relationship between energy and momentum for any particle moving at a speed so close to the speed of light that its total energy E greatly exceeds its rest energy  $m_0c^2$ .

For the high energy particle accelerators we can combine Equations (25) and (32) to get

$$E = pc = qBrc \tag{33}$$

Consider CERN's Super Proton Synchrotron or SPS, shown by the smaller solid circle in Figure (30), which was used to discover the particles responsible for the weak interaction. In this accelerator, the magnets produced fields of B = 1.1 tesla, and the radius of the ring was r = 1.3km (for a circumference of 8km). Thus we have

E = qBrc  
= 
$$(1.6 \times 10^{-19} \text{ coulombs}) \times (1.1 \text{ tesla})$$
  
 $\times (1.3 \times 10^{3} \text{ m}) \times (3 \times 10^{8} \text{ m/s})$   
=  $6.9 \times 10^{-8}$  joules

Converting this answer to electron volts, we get

$$E = \frac{6.9 \times 10^{-8} \text{ joules}}{1.6 \times 10^{-19} \frac{\text{joules}}{\text{eV}}} = 430 \times 10^{9} \text{ eV}$$
  
= 430GeV (34)

How good was our approximation that we could neglect the particle's rest energy and use the simple Equation (32)? Recall that the rest energy of a proton is about 1GeV. Thus the SPS accelerator produced protons with a kinetic energy 430 times greater! For these particles it is not much of an error to neglect the rest energy.

#### Exercise 7

a) The Fermi lab accelerator, with its radius of 1 kilometer, uses superconducting magnets to produce beams of protons with a kinetic energy of 1000 GeV  $(10^{12} \text{eV})$ . How strong a magnetic field is required to produce protons of this energy?

b) Iron electromagnets cannot produce magnetic fields stronger than 2 tesla, which is why superconducting magnets were required to produce the 1000 GeV protons discussed in part a). Before the ring with superconducting magnets was constructed, a ring using iron magnets already existed in the same tunnel. The iron magnets could produce 1.5 tesla fields. What was the maximum energy to which protons could be accelerated before the superconducting magnets were installed? (You can see both rings of magnets in Figure 28.)

#### **Exercise 8**

The large electron-positron (LEP) collider, being constructed at CERN, will create head on collisions between electrons and positrons. (Electrons will go around one way, and positrons, having the opposite charge, will go around the other.) The path of the LEP accelerator, which will have a circumference of 27 km, is shown in Figure (30), superimposed on the countryside north of Geneva, Switzerland.

a) Assuming that the LEP accelerator will use 3 tesla superconducting magnets, what will be the maximum kinetic energy, in eV, of the electrons and positrons that will be accelerated by this machine?

b) What will be the speed of these electrons and positrons? (How many 9's in v/c?)



**Figure 31a** Schematic diagram of the Berkeley 10-inch hydrogen bubble chamber.



#### Figure 31b

The 10-inch Bubble chamber at the Lawrence Radiation Laboratory, University of California, Berkeley. (Photograph copyright The Ealing Corporation, Cambridge, Mass.)

# **BUBBLE CHAMBERS**

In the study of elementary particles, it is just as important to have adequate means of observing particles as it is to have accelerating machines to produce them. One of the more useful devices for this purpose is the bubble chamber invented by Donald Glaser in 1954.

It may not be true that Glaser invented the bubble chamber while looking at the streaks of bubbles in a glass of beer. But the idea is not too far off. When a charged particle like an electron, proton or some exotic elementary particle, passes through a container of liquid hydrogen, the charged particle tends to tear electrons from the hydrogen atoms that it passes, leaving a trail of ionized hydrogen atoms. If the pressure of the liquid hydrogen is suddenly reduced the liquid will start boiling if it has a "seed"—a special location where the boiling can start. The trail of ionized hydrogen atoms left by the charged particle provides a trail of seeds for boiling. The result is a line of bubbles showing where the particle went.

In a typical bubble chamber, a stereoscopic camera is used to record the three dimensional paths of the particles. It is impressive to look at the three dimensional paths in stereoscopic viewers, but unfortunately all we can conveniently do in a book is show a flat two dimensional image like the one in Figure (32). In that picture we see the paths of some of the now more common exotic elementary particles. In the interesting part of this photograph, sketched above, a negative  $\pi^-$  meson collides with a positive proton to create a neutral  $\Lambda^0$  and a neutral K<sup>0</sup> meson. The neutral  $\Lambda^0$  and K<sup>0</sup> do not leave tracks, but they are detected by the fact that the K<sup>0</sup> decayed into a  $\pi^+$  and a  $\pi^-$  meson, and the  $\Lambda^0$  decayed into a  $\pi^-$  and a proton p<sup>+</sup>, all of which are charged particles that left tracks.

To analyze a picture like Figure (32) you need more information than just the tracks left behind by particles. You would also like to know the charge and the momentum or energy of the particles. This is done by placing the bubble chamber in a magnetic field so that positive particle tracks are curved one way and negatives ones the other. And, from Equation (25b), we see that the radii of the tracks tell us the momenta of the particles. Another example of a bubble chamber photograph is Figure (33) where we see the spiral path produced by an electron. The fact that the path is spiral, that the radius of the path is getting smaller, immediately tells us that the electron is losing momentum and therefore energy as it moves through the liquid hydrogen. The magnetic field used for this photograph had a strength B = 1.17 tesla, and the initial radius of the spiral was 7.3 cm. From this we can determine the momentum and energy of the electron.





#### Figure 33

Spiraling electron. An electron enters the chamber at the lower left and spirals to rest as it loses momentum. The spiral track is caused by the magnetic field applied to the chamber which deflects a charged particle into a curved path with a radius of curvature proportional to the particle's momentum. The straight track crossing the spiral is a proton recoiling from a collision with a stray neutron. Because the proton has much greater mass than the electron, its track is much less curved.

## **Exercise 9**

Calculate the energy, in eV of the electron as it entered the photograph in Figure (33). Since you do not know off hand whether the particle was relativistic or not, use the exact relation

$$E^{2} = p^{2}c^{2} + m_{0}^{2}c^{4}$$
 (31)

to determine E from p. From your answer decide whether you could have used the non relativistic formula  $KE = 1/2 m_0 v^2$  or the fully relativistic formula E = pc, or whether you were in an intermediate range where neither approximation works well.



## Figure 32

Bubble chamber photograph showing the creation of a  $K^0$  meson and a  $\Lambda^0$  particle, and their subsequent annihilations. We now know that the K meson is a quark/anti quark pair, and the  $\Lambda^0$  particle contains 3 quarks as does a proton and a neutron. The K and  $\Lambda$  particles last long enough to be seen in a bubble chamber photograph because they each contain a strange quark which decays slowly via the weak interaction. (Photo copyright The Ealing Corporation Cambridge, Mass.)

# **The Mass Spectrometer**

A device commonly seen in chemistry and geology labs is the mass spectrometer which is based on the circular orbits that a charged particle follows in a uniform magnetic field. Figure (34) is a sketch of a mass spectrometer which consists of a semi circular evacuated chamber with a uniform magnetic field  $\vec{B}$ directed up out of the paper. The direction of  $\vec{B}$  is chosen to deflect positive ions around inside the chamber to a photographic plate on the right side. The ions to be studied are boiled off a heated filament and accelerated by a negative cap in a reversed voltage electron gun shown in Figure (35). By measuring the position where the ions strike the photographic plate, we know the radius of the orbit taken by the ion. Combine this with the knowledge of the field B of the spectrometer, and we can determine the ion's momentum p if the charge q is known. The speed of the ion is determined by the accelerating voltage in the gun, thus knowing p gives us the mass m of the ions. Non relativistic formulas work well and the calculations are relachyoinseintiEighner (25).anstes Equations patho 28.th-

Mass spectrometers are used to identify elements in small sample of material, and are particularly useful i being able to separate different isotopes of an element Two different isotopes of an element have differen numbers of neutrons in the nucleus, everything else



#### Figure 34

Top view of a mass spectrograph. A uniform magnetic field  $\vec{B}$  rises directly up through the chamber. The beam of atoms is produced by the accelerating gun shown in Figure 35.

being the same. Thus ions of the two isotopes will have slightly different masses, and land at slightly different distances down the photographic plate. If an isotope is missing in one sample the corresponding line on the photographic plate will be absent. The analogy between looking at the lines identifying isotopes, and looking at a photographic plate showing the spectrum of light, suggested the name *mass spectrometer*.

## **Exercise 10**

Suppose that you wish to measure the mass of an iodine atom using the apparatus of Figures (34) and (35). You coat the filament of the gun in Figure (35) with iodine, and heat the filament until iodine atoms start to boil off. In the process, some of the iodine atoms lose an electron and become positive ions with a charge +e. The ions are then accelerated in the gun by a battery of voltage  $V_{\rm b}$  and then pass into the evacuated chamber.

(a) Assuming that  $V_b = 125$  volts (accelerating voltage) and B = 1000 gauss (0.1 tesla), and that the iodine atoms follow a path of radius r = 18.2 cm, calculate the mass m of the iodine atoms.

(b) How many times more massive is the iodine ion than a proton? From the fact that protons and neutrons have about the same mass, and that an electron is 2000 times lighter, use your result to estimate how many nuclear particles (protons or neutrons) are in an iodine nucleus.



## Figure 35

When the substance to be studied is heated by a filament, atoms evaporate and some lose an electron and become electrically charged positive ions. The ions are then accelerated by an electric field to produce a beam of ions of known kinetic energy.

# **Magnetic Focusing**

In the magnetic force examples we have considered so far, the velocity  $\vec{v}$  of the charged particle started out perpendicular to  $\vec{B}$  and we got the circular orbits we have been discussing.

If we place an electron gun so that the electron beam is aimed down the axis of a pair of Helmholtz coils, as shown in Figure (36), the electron velocity  $\vec{v}$  is parallel to  $\vec{B}$ ,  $\vec{v} \times \vec{B} = 0$  and there is no magnetic force.

Figure (36) is a bit too idealized for the student built electron gun we have been using in earlier examples. Some of the electrons do come out straight as shown in Figure (36), but many come out at an angle as shown in Figure (37a). In Figure (37b) we look at the velocity components  $\vec{v}_{\perp}$  and  $\vec{v}_{\parallel}$  of an electron emerging at an angle q. Because  $\vec{v}_{\parallel} \times \vec{B} = 0$  only the perpendicular component  $\vec{v}_{\perp}$  contributes to the magnetic force

$$\vec{F}_{\text{mag}} = q\vec{v}_{\perp} \times \vec{B}$$
(36)

This force is perpendicular to both  $\vec{v}_{\perp}$  and  $\vec{B}$  as shown in the end view of the electron gun, Figure (37b). In this end view, where we can't see  $\vec{v}_{\parallel}$ , the electron appears to travel around the usual circular path.



#### Figure 36

Electron gun inserted so that the beam of electrons moves parallel to the magnetic field of the coils. If the beam is truly parallel to  $\vec{B}$ , there will be no magnetic force on the electrons.



#### Figure 37a





#### Figure 37b

Consider an electron emerging from the cap at an angle  $\theta$  from the center line as shown in the side view above. Such an electron has a component of velocity  $\vec{v}_{\perp}$  perpendicular to the magnetic field. This produces a magnetic force  $\vec{F}_B = (-e)\vec{v}_{\perp} \times \vec{B}$  which points toward the axis of the gun. The magnetic force  $\vec{F}_B$  can be seen in the end view above. From the end view the electron will appear to travel in a circle about the axis of the gun. The stronger the magnetic field, the smaller the radius of the circle. It is in the side view, Figure (38) that we see the effects of  $\vec{v}_{\parallel}$ . Since there is no force related to  $\vec{v}_{\parallel}$ , this component of velocity is unchanged and simply carries the electron at a constant horizontal speed down the electron gun. The quantity  $\vec{v}_{\parallel}$  is often called the drift speed of the particle. (The situation is not unlike projectile motion, where the horizontal component  $v_x$ of the projectile's velocity is unaffected by the vertical acceleration  $a_y$ .)

When we combine the circular motion, seen in the end view of Figure (37c), with the constant drift speed  $\vec{v}_{||}$ , down the tube seen in Figure (38a) the net effect is a helical path like a stretched spring seen in Figure (38b). The electron in effect spirals around and travels along the magnetic field line. The stronger the magnetic field, the smaller the circle in Figure (37c), and the tighter the helix.



Figure 38a

In the side view of the motion of the electron, we see that  $v_{\parallel}$  is unchanged,  $v_{\parallel}$  just carries the electron down the tube.



helical motion of the electron

## Figure 38b

Oblique view of the helical motion of the electron. When you combine the uniform motion of the electron down the tube with the circular motion around the axis of the tube, you get a helical motion with the same shape as the wire in a stretched spring. The tightening of the helix is seen in Figure (39) where in (a) we see an electron beam with no magnetic field. The electrons are spraying out in a fairly wide cone. In (b) we have a 75 gauss magnetic field aligned parallel to the axis of the gun and we are beginning to see the helical motion of the electrons. In (c) the magnetic field is increased to 200 gauss and the radius of the helix has decreased considerably. As B is increased, the electrons are confined more and more closely to a path along the magnetic field lines. In our electron gun, the magnetic field is having the effect of focusing the electron beam.



a) No magnetic field



b) B = 75 gauss



c) B = 200 gauss





## Figure 39

Focusing an electron beam with a parallel magnetic field. The beam travels along a helical path which becomes tighter as the strength of the magnetic field is increased.

# SPACE PHYSICS

Even in non-uniform magnetic fields there is a tendency for a charged particle to move in a spiral path along a magnetic field line as illustrated in Figure (40). This is true as long as the magnetic field is reasonably uniform over a distance equal to the radius r of the spiral (from Equation (25),  $r = mv_{\perp}/q\vec{v}_B$ ). Neglecting the spiral part of the motion, we see that the large scale effect is that charged particles tend to move or flow along magnetic field lines. This plays an important role in space physics phenomena which deals with charged particles emitted by the sun (the "solar wind") and the interaction of these particles with the magnetic field of the earth and other planets.

There are so many interesting and complex effects in the interaction of the solar wind with planetary magnetic fields that space physics has become an entire field of physics. Seldom are we aware of these effects unless a particularly powerful burst of solar wind particles disrupts radio communications or causes an Aurora Borealis to be seen as far south as the temperate latitudes. The Aurora are caused when particles from the solar wind spiral in along the earth's magnetic field lines and end up striking atoms in the upper atmosphere. The atoms struck by the solar wind particles emit light just like the residual air atoms struck by the electrons in an electron gun.



When charged particles from the sun enter the earth's magnetic field, they spiral around the magnetic field lines much like the electrons in the magnetic focusing experiment of Figure 39.

# The Magnetic Bottle

If a magnetic field has the correct shape, if the field lines pinch together as shown at the left or the right side of Figure (41), then the magnetic force  $\vec{F}_{mag}$  on a charged particle has a component that is directed back from the pinch. For charged particles with the correct speed, this back component of the magnetic force can reflect the particle and reverse  $\vec{v}_{\parallel}$ . If the magnetic field is pinched at both ends, as in Figure (41) the charged particle can reflect back and forth, trapped as if it were in a magnetic bottle.

In the subject of plasma physics, one often deals with hot ionized gasses, particularly in experiments designed to study the possibility of creating controlled fusion reactions. These gases are so hot that they would melt and vaporize any known substance they touch. The only known way to confine these gases to do experiments on them is either do the experiments so fast that the gas does not have time to escape (inertial confinement), or use magnetic fields and devices like the magnetic bottle shown in Figure (41) (magnetic confinement).





Magnetic bottle. When the magnetic field lines pinch together, the charged particles can be reflected back in a process called magnetic mirroring. (At the two ends of the magnetic bottle above, the magnetic force  $\vec{F}_B$  has a component back into the bottle.)

## Van Allen Radiation Belts

The earth's magnetic field shown in Figure (9) and repeated in Figure (42) forms magnetic bottles that can trap charged particles from the solar wind. The ends of the bottles are where the field lines come together at the north and south magnetic poles, and the regions where significant numbers of particles are trapped are called the Van Allen radiation belts shown in Figure (42). Protons are trapped in the inner belt and electrons in the outer one.

It is not feasible to do hand calculations of the motion of charged particles in non-uniform magnetic fields. The motion is just too complicated. But computer calculations, very similar to the orbit calculations discussed in Chapter 4, work well for electric and magnetic forces. As long as we have a formula for the shape of  $\vec{E}$  or  $\vec{B}$ , we can use the Lorentz force law (Equation 20)

$$\vec{F} = q\vec{E} + q\vec{v} \times \vec{B}$$

as one of the steps in the computer program. The computer does not care how complicated the path is, but we might have trouble drawing and interpreting the results.



#### Figure 42

Charged particles, trapped by the earth's magnetic field, spiral around the magnetic field lines reflecting where the lines pinch together at the poles. The earth's magnetic field thus forms a magnetic bottle, holding the charged particles of the Van Allen radiation belts.

In Figure (43), a student, Jeff Lelek, started with the formula for a "dipole magnetic field", namely

$$\vec{\mathbf{B}} = -\left(\frac{\mathbf{B}_0}{\mathbf{R}^3}\right) * \left(\hat{\mathbf{Z}} - 3 * \left(\hat{\mathbf{Z}} \cdot \hat{\mathbf{R}}\right) * \hat{\mathbf{R}}\right)$$
(37)

which is a reasonably accurate representation of the earth's magnetic field, and calculated some electron orbits for this field. The result is fairly complex, but we do get the feeling that the electron is spiraling around the magnetic field lines and reflecting near the magnetic poles.

To provide a simpler interpretation of this motion, the student let the calculation run for a long time, saving up the particle coordinates at many hundreds of different points along the long orbit. These points are then plotted as the dot pattern shown in Figure (43). (In this picture, the latitude of the particle is ignored, the points are all plotted in one plane so we can see the extent of the radial and north-south motion of the particles.) The result gives us a good picture of the distribution of particles in a Van Allen radiation belt. This and similar calculations are discussed in the supplement on computer calculations with the Lorentz force law.



#### Figure 43

Computer plot of the motion of a proton in a dipole magnetic field. The formula for this field and the computer program used to calculate the motion of the proton are given in the Appendix. As you can see, the motion is relatively complex. Not only does the proton reflect back and forth between the poles, but also precesses around the equator.



Figure 44

In this computer plot, all the data points from Figure 43 are plotted as dots in one plane. From this we see the shape of a Van Allen radiation belt emerge. (Figures (43) and (44) from a student project by Jeff Lelek.)

# Chapter 29 Ampere's Law

In this chapter our main focus will be on Ampere's law, a general theorem that allows us to calculate the magnetic fields of simple current distributions in much the same way that Gauss' law allowed us to calculate the electric field of simple charge distributions.

As we use them, Gauss' and Ampere's laws are integral theorems. With Gauss' law we related the total flux out through a closed surface to  $1/\epsilon_0$  times the net charge inside the surface. In general, to calculate the total flux through a surface we have to perform what is called a **surface integral**. Ampere's law will relate the integral of the magnetic field around a closed path to the total currentflowing through that path. This integral around a closed path is called a **line integral**.

Until now we have concentrated on examples that did not require us to say much about integration. But as we discuss Ampere's law in this chapter and the remaining Maxwell equations in the next few chapters, it will be convenient to draw upon the formalism of the surface and line integral. Therefore we will take a short break to discuss the mathematical concepts involved in these integrals.
# THE SURFACE INTEGRAL

In our discussion of Gauss' law near the end of Chapter 24, we defined the flux  $\Phi$  of a fluid in a flow tube as the *amount of water per second flowing past the cross-sectional area of the tube* as shown in Figure (1). This is equal to the velocity v times the cross-sectional area  $A_{\perp}$  of the tube as given in Equation (24-46)

$$\Phi = vA_{\perp} \qquad flux in a flow tube \qquad (24-46)$$

As seen in Figure (2), if we slice the flow tube by a plane that is not normal to the flow tube, the area A of the intersection of the tube and plane is larger than the cross-sectional area  $A_{\perp}$ . The relationship is  $A_{\perp} = A \cos \theta$  where  $\theta$  is the angle between  $\vec{v}$  and  $\vec{A}$ (see Equation 24-45). Defining the vector  $\vec{A}$  as having a magnitude A and direction normal to the plane, we have

$$\vec{v} \cdot \vec{A} = vA\cos\theta = vA_{\perp}$$

and the formula for the flux in the flow tube is

$$\Phi = \vec{v} \cdot \vec{A} \tag{24-47a}$$

In Chapter 24 we considered only problems where  $A_{\perp}$  was something simple like a sphere around a point source, or a cylinder around a line source, and we could easily write a formula for the total flux. We now wish to consider how we should calculate, at least in principle, the flux in a more complex flow like the stream shown in Figure (3).

To give our flux calculation a sense of reality, suppose that we wish to catch all the salmon swimming up a stream to spawn. As shown in Figure (3), we place a net



### Figure 1

The flux of water  $\boldsymbol{\Phi}$  through a flow tube is the amount of water per second flowing past a cross-sectional area  $A_1$ .

that goes completely across the stream, from bank to bank, from the surface to the bottom. The total flux  $\Phi_T$  of water flowing through this net is therefore equal to the total current in the stream.

To calculate the total flux  $\Phi_T$ , we break the stream flow up into a number of small flow tubes bounded by stream lines as shown in (3). Focusing our attention on the i th flow tube, we see that the tube intersects an area  $d\vec{A}_i$  of the fish net. The flux through the fish net due to the i th tube is

$$d\Phi_i = \vec{v}_i \cdot d\vec{A}_i \tag{1}$$

where  $\vec{v}_i$  is the velocity of the water at the intersection of the tube and the net. The total flux or current of water in the stream is simply the sum of the fluxes in each flow tube, which can be written

$$\Phi_{\rm T} = \sum_{\substack{\text{all flow} \\ \text{tubes}}} \Phi_{\rm i} = \sum_{\rm i} \vec{\rm v}_{\rm i} \cdot d\vec{\rm A}_{\rm i}$$
(2)

where the  $d\vec{A}_i$  are just those areas on the fish net marked out by the flow tubes.

If we go to infinitesimal sized flow tubes, the sum in Equation (2) becomes an integral which can be written as



### Figure 2

If we have an area A that is not normal to the stream, then the cross-sectional area is  $A_{\perp} = A \cos \theta$ , and the flux is  $\Phi = vA_{\perp} = vA \cos \theta$ , which can also be written  $\Phi = \vec{v} \cdot \vec{A}$ . where the  $d\vec{A}$ 's are infinitesimal pieces of area on the fish net and our sum or integration extends over the entire submerged area of the net. Because we are integrating over an area or surface in Equation (3), this integral is called a *surface integral*.

Think of Equation (3), not as an integral you "do", like  $\int x^2 dx = x^3/3$ , but more as a formal statement of the steps we went through to calculate the total flux  $\Phi_T$ .

Suppose, for example, someone came up to you and asked how you would calculate the total current in the stream. If you were a mathematician you might answer, "I would calculate the integral

$$\Phi_{\rm T} = \int_{\rm S} \vec{\rm v} \cdot d\vec{\rm A} \tag{3a}$$

where S is a surface cutting the stream."

If you were a physicist, you might answer, "Throw a fish net across the stream, making sure that there are no gaps that the fish can get through. (This defines the mathematician's surface S). Then measure the flux of water through each hole in the net (these are the  $\vec{v} \cdot d\vec{A}$ 's of Equation 3a), and then add them up to get the total flux (do the integral)."

Basically, the mathematician's statement in Equation (3a) is short hand notation for all the steps that the physicist would carry out.



### Figure 3

To calculate the flux of water through a fish net, we can first calculate the flux of water through each hole in the net, and then add up the fluxes to get the total flux.

## **Gauss' Law**

The statement of Gauss' law applied to electric fields in Chapter (24) was that the total electric flux  $\Phi_T$  out through a closed surface was equal to  $1/\epsilon_0$  times the total charge Q<sub>in</sub> inside the surface. Our surface integral of Equation (3) allows us to give a more formal (at least more mathematical sounding) statement of Gauss' law.

Suppose we have a collection of charged particles as shown in Figure (4), which are completely surrounded by a closed surface S. (*Think of the closed surface as being the surface of an inflated balloon. There cannot be any holes in the surface or air would escape.*) The total flux of the field E out through the surface S is formally given by the surface integral

$$\Phi_{\rm T} = \int_{\rm S} \vec{\rm E} \cdot d\vec{\rm A} \tag{4}$$

where the  $d\vec{A}$  are small pieces of the surface, and  $\vec{E}$  is the electric field vector at each  $d\vec{A}$ .



## Figure 4

Closed surface S completely surrounding a collection of charges. The flux of  $\vec{E}$  out through the closed surface is equal to  $1/\varepsilon_0$  times the total charge inside.

The total charge  $Q_{in}$  inside the surface S is obtained by adding up all the charges we find inside. Any charges outside do not count. (We have to have a completely closed surface so that we can decide whether a charge is inside or not.) Then equating the total flux  $\Phi_T$  to  $Q_{in}/\epsilon_0$  we get the *integral equation* 

There is nothing really new in Equation (5) that we did not say back in Chapter (24). What we now have is a convenient short hand notation for all the steps we discussed earlier.

We will now use Equation (5) to calculate the electric field of a point charge. Although we have done this same calculation before, we will do it again to remind us of the steps we actually go through to apply Equation (5). A formal equation like this becomes real or useful only when we have an explicit example to remind us how it is used. *When you memorize such an equation, also memorize an example to go with it.* 

In Figure (5), we have a point charge +Q that produces a radial electric field  $\vec{E}$  as shown. To apply Gauss' law we draw a spherical surface S of radius r around the charge. For this surface we have

$$\Phi_{\rm T} = \int_{\rm S} \vec{\rm E} \cdot d\vec{\rm A} = {\rm E}{\rm A}_{\perp} = {\rm E}({\rm r})4\pi{\rm r}^2$$

Inside the surface the total charge is Q. Thus Gauss' law, Equation (5), gives

$$\int_{S} \vec{E} \cdot d\vec{A} = \frac{Q_{in}}{\varepsilon_{0}}$$

$$E(r)4\pi r^{2} = \frac{Q}{\varepsilon_{0}}$$

$$E(r) = \frac{Q}{4\pi\varepsilon_{0}r^{2}}$$
(6)

The only thing that is new here is the use of the notation  $\int_{S} \vec{E} \cdot d\vec{A}$  for total flux  $\Phi_{T}$ . When we actually wish to calculate  $\Phi_{T}$ , we look for a surface that is perpendicular to  $\vec{E}$  so that we can use the simple formula  $EA_{\parallel}$ .

If the charge distribution were complex, more like Figure (4), we could calculate  $\int_{S} \vec{E} \cdot d\vec{A}$  by casting a fish net all around the charges and evaluating  $\vec{E}_i \cdot d\vec{A}_i$  for each hole in the net. The formal expression of the surface integral at least gives us a procedure we can follow if we are desperate.



### Figure 5

For the electric field of a point charge, we know immediately that the total flux  $\Phi_T$  out through the spherical surface is the area  $4\pi r^2$  times the strength  $\vec{E}(r)$  of the field. Thus

$$\boldsymbol{\varPhi}_{T} \equiv \int\limits_{S} \vec{E} \cdot d\vec{A} = \vec{E}(r) \times 4\pi r^{2}$$

# THE LINE INTEGRAL

Another formal concept which we will use extensively in the remaining chapters on electromagnetic theory is the line integral. You have already been exposed to the idea in earlier discussions of the concept of work. If we exert a force  $\vec{F}$  on a particle while the particle moves from Point (1) to Point (2) as shown in Figure (6), then the work we do is given by the integral

Work W = 
$$\int_{1}^{2} \vec{F} \cdot d\vec{x}$$
 (7)

where we are integrating along the path in Figure (6).

Equation (7) is short hand notation for many steps. What it really says is to draw the path taken by the particle in going from Point (1) to (2), break the path up into lots of little steps  $d\vec{x}_i$ , calculate the work  $dW_i$  we do during each step,  $dW_i = \vec{F}_i \cdot d\vec{x}_i$ , and then add up all the  $dW_i$  to get the total work W.



To calculate the total work done moving a particle from point (1) to point (2) along a path, first break the path up into many short displacements  $d\vec{x}$ . The work  $dW_i$  is  $dW_i = \vec{F}_i \cdot d\vec{x}_i$ . The total work W is the sum of all the  $dW_i$ . The first thing we have to worry about in discussing Equation (7), is what path the particle takes in going from Point (1) to Point (2). If we are moving an eraser over a blackboard, the longer the path, the more work we do. In this case, we cannot do the line integral until the path has been specified.

On the other hand, if we are carrying the particle around the room, exerting a force  $\vec{F} = -\vec{F}_g$  that just overcomes the gravitational force, then the work we do is stored as gravitational potential energy. The change in potential energy, and therefore the line integral of Equation (7), depends only on the end points (1) and (2) and not on the path we take. When the line integral of a force does not depend upon the path, we say that the force is **conservative**.

A formal statement that a force is conservative is that the line integrals are equal for any two paths -- for example, path (a) and path (b) in Figure (7).

$$\int_{1(\text{path a})}^{2} \vec{F} \cdot d\vec{x} = \int_{1(\text{path b})}^{2} \vec{F} \cdot d\vec{x}$$
(8)



### Figure 7

If the work done in carrying the particle from point (1) to point (2) does not depend upon which path we take, we say that the force is conservative.

Let us write Equation (8) in the form

$$\int_{1(\text{path a})}^{2} \vec{F} \cdot d\vec{x} - \int_{1(\text{path b})}^{2} \vec{F} \cdot d\vec{x} = 0$$

Now take the minus sign inside the integral over Path (b) so that we have a sum of  $\vec{F}_{i} \cdot (-d\vec{x}_{i})$ 

$$\int_{1(\text{path a})}^{2} \vec{F} \cdot d\vec{x} + \int_{1(\text{path b})}^{2} \vec{F} \cdot \left(-d\vec{x}\right) = 0$$
(9)

For the path (b) integral, we have reversed the direction of each step. The sum of the reversed steps is the same as going back, from point (2) to point (1) as illustrated in Figure (8). Thus Equation (9) becomes

$$\int_{1(\text{path a})}^{2} \vec{F} \cdot d\vec{x} + \int_{2(\text{path b})}^{1} \vec{F} \cdot d\vec{x} = 0$$
(10)

If Equation (10) applies for any Path (a) and (b), then the force  $\vec{F}$  is conservative.



### Figure 8

If we take path b backwards, i.e. go from point (2) to point (1), the  $d\vec{x}_i$  on path b are reversed and the integral along path b changes sign. If the line integral from (1) to (2) does not depend upon the path, then the line integral for any return trip must be the negative of the integral for the trip out, and the sum of the two integrals must be zero.

Equation (10) does not really depend upon the Points (1) and (2). More generally, it says that if you go out and then come back to your starting point, and the sum of all your  $\vec{F}_i \cdot d\vec{x}_i$  is zero, then the force is conservative. This special case of a line integral that comes back to the starting point as in Figure (9) is called *the line integral around a closed path*, and is denoted by an integral sign with a circle in the center

$$\oint \vec{F} \cdot d\vec{x} \equiv \begin{cases} \text{the line integral} \\ \text{around a closed path} \\ \text{as in Figure 9} \end{cases}$$
(11)

With the notation of Equation (11), we can formally define a conservative force  $\vec{F}$  as one for which

$$\oint_{\substack{\text{for any}\\ \text{closed path}}} \vec{F} \cdot d\vec{x} = 0$$

$$definition of a (12)$$

This line integral around a closed path will turn out to be an extremely useful mathematical tool. We have already seen that it distinguishes a conservative force like gravity, where  $\oint \vec{F} \cdot d\vec{x} = 0$ , from a non conservative force like friction on a blackboard eraser, for which  $\oint \vec{F} \cdot d\vec{x} \neq 0$ . In another case, namely Ampere's law to be discussed next, the line integral of the magnetic field around a closed path tells us something about the currents that flow through the path.





For a conservative force, the line integral  $\oint \vec{F} \cdot d\vec{x}$ , that goes completely around a closed path, must be zero. It is not necessary to specify where the calculation starts.

## AMPERE'S LAW

ar

Figure (10), which is similar to Figure (28-14), is a sketch of the magnetic field produced by a current in a straight wire. In this figure the current i is directed up and out of the paper, and the magnetic field lines travel in counter clockwise circles as shown. We saw from equation (28-18) that the strength of the magnetic field is given by

$$\vec{\mathbf{B}} = \frac{\mu_0 \mathbf{i}}{2\pi \mathbf{r}} \tag{28-18}$$

In Figure (11) we have drawn a circular path of radius r around the wire and broken the path into a series of steps indicated by short vectors  $d\vec{\ell}_i$ . We drew the stick figure to emphasize the idea that this is really a path and that  $d\vec{\ell}_i$  shows the length and direction of the i th step.

For each of the steps, calculate the dot product  $\vec{B}_i \cdot d\vec{\ell}_i$ where  $\vec{B}_i$  is the magnetic field at that step, and then add up the  $\vec{B}_i \cdot d\vec{\ell}_i$  for all the steps around the path to get

$$\sum_{\substack{\text{all steps}\\\text{ound path}}} \vec{B}_i \cdot d\vec{\ell}_i \to \oint \vec{B} \cdot d\vec{\ell}$$
(13)

The result is the line integral of  $\vec{B}$  around the closed path.

Why bother calculating this line integral? Let us put in the value for  $|\vec{B}|$  given by Equation (28-18) and see why. We happen to have chosen a path where each step  $d\vec{\ell}_i$  is parallel to  $\vec{B}$  at that point, so that

$$\vec{\mathbf{B}}_{i} \cdot d\vec{\ell}_{i} = \mathbf{B}_{i} d\ell_{i} \tag{14}$$



Circular magnetic field of a wire.

and Equation (13) becomes

$$\oint \vec{B} \cdot d\vec{\ell} = \oint B d\ell \tag{15}$$

In addition, our path has a constant radius r, so that  $B = \mu_0 i/2\pi r$  is constant all around the path. We can take this constant outside the integral in Equation (15) to get

$$\oint \vec{\mathbf{B}} \cdot d\vec{\ell} = \mathbf{B} \oint d\ell \tag{16}$$

Next we note that  $\oint d\ell$  is just the sum of the lengths of our steps around the circle; i.e., it is just the circumference  $2\pi r$  of the circle, and we get

$$\oint \vec{B} \cdot d\vec{\ell} = B \oint d\ell = B \times 2\pi r$$
 (17)

Finally substituting the value of B from Equation (28-18) we get



### Figure 11

Circular path of radius r around the wire. As we walk around the path, each step represents a displacement  $d\dot{\ell}$ . To calculate the line integral  $\oint \vec{B} \cdot d\hat{\ell}$ , we take the dot product of  $d\hat{\ell}$  with B at each interval and add them up as we go around the entire path. In this case the result is simply B(r) times the circumference  $2\pi r$  of the path. There are several points we want to make about Equation (18). First we made the calculation easy by choosing a circular path that was parallel to  $\vec{B}$  all the way around. This allowed us to replace the dot product  $\vec{B} \cdot d\vec{\ell}$  by a numerical product  $Bd\ell$ , pull the constant B outside the integral, and get an answer almost by inspection. This should be reminiscent of our work with Gauss' law where we chose surfaces that made it easy to solve the problem.

The second point is that we get an exceptionally simple answer for the line integral of  $\vec{B}$  around the wire, namely

$$\oint \vec{\mathbf{B}} \cdot d\vec{\ell} = \mu_0 \mathbf{i}$$
(18a)

The line integral depends only on the current i through the path and not on the radius r of the circular path.

What about more general paths that go around the wire? To find out, we have to do a slightly harder calculation, but the answer is interesting enough to justify the effort.

In Figure (12) we have constructed a closed path made up of three arc sections of lengths  $r_1\theta_1$ ,  $r_2\theta_2$ , and  $r_3\theta_3$ connected by radial sections as shown. These arcs are sections of circles of radii  $r_1$ ,  $r_2$  and  $r_3$ , respectively. We wish to calculate  $\oint \vec{B} \cdot d\vec{\ell}$  for this path and see how the answer compares with what we got for the circular path.

The first thing to note as we go around our new path is that in all the *radial* sections,  $\vec{B}$  and  $d\vec{\ell}$  are perpendicular to each other, so that  $\vec{B} \cdot d\vec{\ell} = 0$ . The radial sections do not contribute to our line integral and all we have to do is add up the contributions from the three arc segments. These are easy to calculate because  $\vec{B} \cdot d\vec{\ell} = Bd\ell$  and B is constant over each arc, so that the integral of  $\vec{B} \cdot d\vec{\ell}$  over an arc segment is just the value of B times the length  $r\theta$  of the arc. We get

$$\int_{\text{arc 1}} \vec{B} \cdot d\ell = B_1 r_1 \theta_1 = \frac{\mu_0 i}{2\pi r_1} r_1 \theta_1 = \frac{\mu_0 i \theta_1}{2\pi}$$
$$\int_{\text{arc 2}} \vec{B} \cdot d\ell = B_2 r_2 \theta_2 = \frac{\mu_0 i}{2\pi r_2} r_2 \theta_2 = \frac{\mu_0 i \theta_2}{2\pi}$$
$$\int_{\text{arc 3}} \vec{B} \cdot d\ell = B_3 r_3 \theta_3 = \frac{\mu_0 i}{2\pi r_3} r_3 \theta_3 = \frac{\mu_0 i \theta_3}{2\pi}$$

Adding the contribution from each arc segment we get the line integral around the closed path

$$\oint \vec{B} \cdot d\vec{\ell} = \frac{\mu_0 i\theta_1}{2\pi} + \frac{\mu_0 i\theta_2}{2\pi} + \frac{\mu_0 i\theta_3}{2\pi}$$
$$= \frac{\mu_0 i}{2\pi} (\theta_1 + \theta_2 + \theta_3)$$

But  $(\theta_1 + \theta_2 + \theta_3)$  is the sum of the angles around the circle, and is therefore equal to  $2\pi$ . Thus we get for the path of Figure (12)



### Figure 12

A somewhat arbitrary path around the wire is made of arc sections connected by radial sections. Since B is perpendicular to  $d\hat{\ell}$  in the radial sections, the radial sections do not contribute to the  $\oint \vec{B} \cdot d\hat{\ell}$  for this path. In the arcs, the length of the arc increases with r, but B decreases as 1/r, so that the contribution of the arc does not depend upon how far out it is. As a result the  $\oint \vec{B} \cdot d\hat{\ell}$  is the same for this path as for a circular path centered on the wire.

$$\oint \vec{\mathbf{B}} \cdot d\vec{\ell} = \frac{\mu_0 i}{2\pi} (2\pi) = \mu_0 i$$
(19)

which is the same answer we got for the circular path.

The result in Equation (19) did not depend upon how many line segments we used, because each arc contributed an angle  $\theta$ , and if the path goes all the way around, the angles always add up to  $2\pi$ . In Figure (13) we have imitated a smooth path (the dotted line) by a path consisting of many arc sections. The more arcs we use the closer the imitation. We can come arbitrarily close to the desired path using paths whose integral  $\oint \vec{B} \cdot d\vec{\ell}$ is  $\mu_0 i$  In this sense we have proved that Equation (19) applies to any closed path around the wire.

It is another story if the path does not go around the wire. In Figure (14) we have such a path made up of two arc and two radial segments as shown. As before, we can ignore the radial segments because  $\vec{B}$  and  $d\vec{\ell}$  are perpendicular and  $\vec{B} \cdot d\vec{\ell} = 0$ .



#### Figure 13

We can approximate an arbitrary path (dotted line) by a series of connected radial and arc sections. The smaller the angle  $d\theta$  marking the arc sections, the better the approximation. In calculating  $\vec{B} \cdot d\vec{\ell}$ , the radial sections do not count, and we can bring all the arc sections back to a single circle centered on the wire. As a result,  $\oint \vec{B} \cdot d\vec{\ell}$  does not depend upon the shape of the path, as long as the path goes around the wire.

On the outer segment we are going in the same direction as  $\vec{B}$  so that  $\vec{B} \cdot d\vec{\ell}$  is positive and we get

$$\int_{\text{arc 1}} \vec{\mathbf{B}} \cdot d\vec{\ell} = \mathbf{B}_1(\mathbf{r}_1 \mathbf{\theta}) = \frac{\mu_0 \mathbf{i}}{2\pi \mathbf{r}_1} \mathbf{r}_1 \mathbf{\theta} = \frac{\mu_0 \mathbf{i}}{2\pi} \mathbf{\theta}$$

On the inner arc we are coming back around in a direction opposite to  $\vec{B}$ , the quantity  $\vec{B} \cdot d\vec{\ell}$  is negative, and we get

$$\int_{\operatorname{arc} 2} \vec{\mathbf{B}} \cdot d\vec{\ell} = -\mathbf{B}_2(\mathbf{r}_2 \theta) = \frac{-\mu_0 i}{2\pi \mathbf{r}_2} \mathbf{r}_2 \theta = \frac{-\mu_0 i}{2\pi} \theta$$

Adding up the two contributions from the two arcs, we get

$$\oint_{\substack{\text{Path of}\\\text{Figure 14}}} \vec{B} \cdot d\vec{\ell} = \left(\frac{\mu_0 i\theta}{2\pi}\right)_{\text{arc 1}} + \left(\frac{-\mu_0 i\theta}{2\pi}\right)_{\text{arc 2}} = 0 \quad (20)$$

For this closed path which does not go around the current, we get  $\oint \vec{B} \cdot d\vec{\ell} = 0$ . This result is not changed if we add more arcs and radial segments to the path. As long as the path does not go around the current, we get zero for  $\oint \vec{B} \cdot d\vec{\ell}$ .



### Figure 14

In this example, where the path does not go around the wire, the sections labeled  $r_1\theta$  and  $r_2\theta$  contribute equal and opposite amounts to the line integral  $\oint \vec{B} \cdot d\vec{\ell}$ . As a result  $\oint \vec{B} \cdot d\vec{\ell}$  is zero for this, or any path that does not go around the wire.

## **Several Wires**

It is relatively straightforward to generalize our results to the case where we have several wires as in Figures (15 a,b). Here we have three currents  $i_1$ ,  $i_2$ , and  $i_3$  each alone producing a magnetic field  $\vec{B}_1$ ,  $\vec{B}_2$  and  $\vec{B}_3$ respectively.

The first step is to show that the net field  $\vec{B}$  at any point is the vector sum of the fields of the individual wires. We can do this by considering the force on a test particle of charge q moving with a velocity  $\vec{v}$  as shown in Figure (15a). Our earlier results tell us that the current  $i_1$  exerts a force

$$\vec{F}_1 = q\vec{v} \times \vec{B}_1$$

Similarly i2 and i3 exert forces

$$\vec{F}_2 = q\vec{v} \times \vec{B}_2$$
$$\vec{F}_3 = q\vec{v} \times \vec{B}_3$$

Newton's second law required us to take the vector sum of the individual forces to get the total force  $\vec{F}$  acting on an object

$$\vec{F} = \vec{F}_1 + \vec{F}_2 + \vec{F}_3 = q\vec{v} \times (\vec{B}_1 + \vec{B}_2 + \vec{B}_3)$$
(21)

If we write this total force in the form



### Figure 15a

A charge q moving in the vicinity of three currents  $i_1$ ,  $i_2$  and  $i_3$ . If the magnetic field  $\vec{B}$  at the charge is the vector sum of the fields  $\vec{B}_1$ ,  $\vec{B}_2$  and  $\vec{B}_3$  of the three wires, then the net magnetic force  $\vec{F}_B$  on q is given by

$$\vec{F}_B = q\vec{v} \times \vec{B} = q\vec{v} \times \left(\vec{B}_1 + \vec{B}_2 + \vec{B}_3\right)$$
$$= q\vec{v} \times \vec{B}_1 + q\vec{v} \times \vec{B}_2 + q\vec{v} \times \vec{B}_3$$
$$= \vec{F}_{B1} + \vec{F}_{B2} + \vec{F}_{B3}$$

and we get the desired result that the net force on q is the vector sum of the forces exerted by each wire.

$$\vec{F}_{T} = q\vec{v} \times \vec{B}$$
(22)

where  $\vec{B}$  is the effective field acting on the test particle, then Equations (21) and (22) give

$$\vec{B} = \vec{B}_1 + \vec{B}_2 + \vec{B}_3$$
 (23)

The fact that magnetic fields add vectorially is a consequence of the vector addition of forces and our use of the magnetic force law to define  $\vec{B}$ .

With Equation (23) we can now calculate  $\oint \vec{B} \cdot d\vec{l}$  for the field of several wires. Let us draw a path around two of the wires, as shown in Figure (15b). For this path, we get

$$\oint_{\substack{\text{Closed path} \\ \text{of Fig. 29-15}}} \vec{B} \cdot d\vec{\ell} = \oint (\vec{B} \cdot \vec{I} + \vec{B} \cdot \vec{B} \cdot \vec{I}) \cdot d\vec{\ell}$$

$$= \oint \vec{B}_1 \cdot d\vec{\ell} + \oint \vec{B}_2 \cdot d\vec{\ell} + \oint \vec{B}_3 \cdot d\vec{\ell}$$
(24)

Since the closed path goes around currents  $i_1$  and  $i_2$ , we get from Equation (19)



**Figure 15b** Calculating  $\oint \vec{B} \cdot d\vec{l}$  for a path that goes around two of the wires.

$$\oint \vec{B}_1 \cdot d\vec{\ell} = \mu_0 i_1$$
$$\oint \vec{B}_2 \cdot d\vec{\ell} = \mu_0 i_2$$

Since the path misses  $i_3$ , we get

$$\oint \vec{B}_3 \cdot d\vec{\ell} = 0$$

and Equation (24) gives

$$\oint_{\substack{\text{Closed path} \\ \text{of Fig. 15}}} \vec{B} \cdot d\vec{\ell} = \mu_0 (i_1 + i_2) = \mu_0 \times \begin{pmatrix} \text{current} \\ \text{enclosed} \\ \text{by path} \end{pmatrix}$$
(25)

Equation (25) tells us that  $\oint \vec{B} \cdot d\vec{\ell}$  around a closed path is equal to  $\mu_0$  times the total current  $i = i_1 + i_2$  encircled by the path. This has the flavor of Gauss' law which said that the total flux or surface integral of  $\vec{E}$  out through a closed surface was  $1/\epsilon_0$  times the total charge  $Q_{in}$  inside the surface. Just as charge outside the closed surface did not contribute to the surface integral of  $\vec{E}$ , currents outside the closed path do not contribute to the line integral of  $\vec{B}$ .

We derived Equation (25) for the case that all our currents were in parallel straight wires. It turns out that it does not matter if the wires are straight, bent, or form a hideous tangle. As a general rule, if we construct a closed path, then the line integral of  $\vec{B}$  around the closed path is  $\mu_0$  times the net current  $i_{enclosed}$  flowing through the path

$$\oint \vec{B} \cdot d\vec{\ell} = \mu_0 i_{\text{enclosed}}$$

$$any closed
path$$
(26)
$$Ampere's
Law$$

This extremely powerful and general theorem is known as *Ampere's law*.

So far in this chapter we have focused on mathematical concepts. Let us now work out some practical applications of Ampere's law to get a feeling for how the law is used.

## **Field of a Straight Wire**

Our first application of Ampere's law will be to calculate the magnetic field of a straight wire. We will use this trivial example to illustrate the steps used in applying Ampere's law.

First we sketch the situation as in Figure (16), and then write down Ampere's law to remind us of the law we are using

$$\oint \vec{B} \cdot d\vec{\ell} = \mu_0 i_{enclosed}$$

Next we choose a closed path that makes the line integral as simple as possible. Generally the path should either be along  $\vec{B}$  so that  $\vec{B} \cdot d\vec{\ell} = Bd\ell$ , or perpendicular so that  $\vec{B} \cdot d\vec{\ell} = 0$ . The circular path of Figure (16) gives  $\vec{B} \cdot d\vec{\ell} = Bd\ell$  with B constant, thus

$$\oint \vec{B} \cdot d\vec{\ell} = \oint Bd\ell = B \oint d\ell = B * 2\pi r = \mu_0 i$$

The result is

$$B = \frac{\mu_0 i}{2\pi r}$$

which we expected. When you memorize Ampere's law, *memorize an example like this to go with it*.





Using Ampere's law to calculate the magnetic field of a wire. We have  $\oint \vec{B} \cdot d\vec{\ell} = B \times 2\pi r$  around the path. Thus Ampere's law  $\oint \vec{B} \cdot d\vec{\ell} = \mu_0 i$  gives  $B = \mu_0 i/2\pi r$ .

## Exercise 1

Each of the indicated eight conductors in Figure (17) carries 2.0A of current into (dark) or out of (white) the page. Two paths are indicated for the line integral  $\oint \vec{B} \cdot d\vec{\ell}$ . What is the value of the integral for (a) the dotted path? (b) the dashed path?





## **Exercise 2**

Eight wires cut the page perpendicularly at the points shown in Figure (18). A wire labeled with the integer k (k = 1, 2..., 8) bears the current  $ki_0$ . For those with odd k, the current is up, out of the page; for those with even k it is down, into the page. Evaluate  $\oint \vec{B} \cdot d\vec{\ell}$  along the closed path shown, in the direction shown.





## **Exercise 3**

Show that a uniform magnetic field B cannot drop abruptly to zero as one moves at right angles to it, as suggested by the horizontal arrow through point a in Figure (19). (Hint: Apply Ampere's law to the rectangular path shown by the dashed lines.) In actual magnets "fringing" of the lines of B always occurs, which means that B approaches zero gradually.



Figure 19

## **Exercise 4**

Figure (20) shows a cross section of a long cylindrical conductor of radius a, carrying a uniformly distributed current i. Assume a = 2.0 cm, i = 100A, and sketch a plot of B(r) over the range 0 < r < 4 cm.



### Figure 20

(The above are some choice problems from Halliday and Resnick.)

## **Exercise 5**

Figure (21) shows a cross section of a hollow cylindrical conductor of radii a and b, carrying a uniformly distributed current i.

*a*) Show that B(r) for the range b < r < a is given by

B(r) = 
$$\frac{\mu_0 i}{2\pi r} \left( \frac{r^2 - b^2}{a^2 - b^2} \right)$$

b) Test this formula for the special cases of r = a, r = b, and r = 0.

c) Assume a = 2.0 cm, b = 1.8 cm, and i = 100 A. What is the value of B at r = a? (Give your answer in tesla and gauss.)





## **Exercise 6**

Figure (22) shows a cross section of a long conductor of a type called a coaxial cable. Its radii (a, b, c) are shown in the figure. Equal but opposite currents i exist in the two conductors. Derive expressions for B(r) in the ranges

c) b < r < a, and

*d*) r > a.

*e)* Test these expressions for all the special cases that occur to you.



### Figure 22

Exercise 6 is a model of a *coaxial cable*, where the current goes one way on the inner conductor and back the other way on the outside shield. If we draw any circuit outside the cable, there is no net current through the circuit, thus there is no magnetic field outside. As a result, coaxial cables confine all magnetic fields to the inside of the cable. This is important in many electronics applications where you do not want fields to radiate out from your wires. The cables we use in the lab, the ones with the so called BNC connectors, are coaxial cables, as are the cables that carry cable television.

# FIELD OF A SOLENOID

As with Gauss' law, Ampere's law is most useful when we already know the field structure and wish to calculate the strength of the field. The classic example to which ampere's law is applied is the calculation of the magnetic field of a long straight solenoid.

A long solenoid is a coil of wire in which the length L of the coil is considerably larger than the diameter d of the individual turns. The shape of the field produced when a current i flows through the coil was illustrated in Figure (28-21) and is sketched here in Figure (23). Iron filings gave us the shape of the field and Ampere's law will tell us the strength.

The important and useful feature of a solenoid is that we have a nearly uniform magnetic field inside the coil and nearly zero field outside. The longer the solenoid, relative to the diameter d, the more uniform the field  $\vec{B}$  inside and the more nearly it is zero outside.



## Figure 23

Calculating the magnetic field of a long solenoid. Around the path starting at point (1) we have  $\oint \vec{B} \cdot d\vec{\ell} = 0 + Bh + 0 + 0$ . The amount of current enclosed by the path is  $i_{tot} = (nh)i$  where n is the number of turns per unit length. Thus Ampere's law  $\oint \vec{B} \cdot d\vec{\ell} = \mu_0 i_{tot}$  gives  $Bh = \mu_0 nhi$  or  $B = \mu_0 ni$ .

# **Right Hand Rule for Solenoids**

The direction of the field inside the solenoid is a bit tricky to figure out. As shown in Figure (24), up near the wires and in between the turns, the field goes in a circle around the wire just as it does for a straight wire. As we go out from the wire the circular patterns merge to create the uniform field in the center of the solenoid.

We see, from Figure (24), that if the current goes around the coil in such a way that the current is up out of the paper on the right side and down into the paper on the left, then the field close to the wires will go in counterclockwise circles on the right and clockwise circles on the left. For both these sets of circles, the field inside the coil points down. As a result the uniform field inside the coil is down as shown.

There is a simple way to remember this result without having to look at the field close to the wires. Curl the fingers of your *right* hand in the direction of the flow of the current i in the solenoid, and your thumb will point in the direction of the magnetic field inside the solenoid. We will call this the right hand rule for solenoids.



## Figure 24

If you know the direction of the current in the wire, you can determine the direction of the magnetic field by looking very close to the wire where the field goes around the wire. You get the same answer if you curl the fingers of your right hand around in the direction the current in the coil is flowing. Your thumb then points in the direction of the field.

# Evaluation of the Line Integral

Figure (25) is a detail showing the path we are going to use to evaluate  $\oint \mathbf{B} \cdot \mathbf{d} \, \ell$  for the solenoid. This path goes down the solenoid in the direction of B (side 1), and out through the coil (side 2), up where B = 0 (side 3) and back into the coil (side 4).

We can write  $\oint \vec{B} \cdot d\vec{\ell}$  as the sum of four terms for the four sides

$$\oint \vec{B} \cdot d\vec{\ell} = \int_{\text{side 1}} \vec{B} \cdot d\vec{\ell} + \int_{\text{side 2}} \vec{B} \cdot d\vec{\ell} + \int_{\text{side 3}} \vec{B} \cdot d\vec{\ell} + \int_{\text{side 4}} \vec{B} \cdot d\vec{\ell}$$

On sides 2 and 4, when the path is inside the coil,  $\vec{B}$  and  $d\vec{\ell}$  are perpendicular and we get  $\vec{B} \cdot d\vec{\ell} = 0$ . Outside the coil it is still 0 because there is no field there. Likewise  $\mathbf{B} \cdot \mathbf{d} \mathbf{\ell} = 0$  for side 3 because there is no field there. The only contribution we get is from side 1 inside the coil. If h is the height of our path, then

$$\oint \vec{B} \cdot d\vec{\ell} = \int_{\text{side } 1} \vec{B} \cdot d\vec{\ell} = Bh$$
 (27)



## Figure 25

Right hand rule for using Ampere's law. We define the positive direction around the path as the direction you curl the fingers of your right hand when the thumb is pointing in the direction of the current through the path. (As you see, we can come up with a right hand rule for almost anything.)

# **Calculation of** ienclosed

From Figure (25) we see that we get a current i up through our path each time another turn comes up through the path. On the left side of the coil the current goes down into the paper, but these downward currents lie outside our path and therefore are not included in our evaluation of  $i_{enclosed}$ . Only the positive upward currents count, and  $i_{enclosed}$  is simply i times the number of turns that go up through the path.

To calculate the number of turns in a height h of the coil, we note that if the coil has a length L and a total of N turns, then the number of turns per unit length n is given by

$$\begin{array}{c} \text{number of turns} \\ \text{per unit length} \end{array} \right\} \equiv n = \frac{N}{L}$$
 (28)

and in a height h there must be nh turns

$$\begin{array}{c} \text{number of turns} \\ \text{in a height h} \end{array} = nh \tag{29}$$

With nh turns, each carrying a current i, going up through our path, we see that  $i_{enclosed}$  must be

$$i_{\text{enclosed}} = \text{inh}$$
 (30)

## **Using Ampere's law**

We are now ready to apply Ampere's law to evaluate the strength B of the field inside the solenoid. Using Equation (27) for  $\oint \vec{B} \cdot d\vec{\ell}$ , and Equation (31) for i<sub>enclosed</sub>, we get

$$\int \vec{B} \cdot d\vec{\ell} = \mu_0 i_{\text{enclosed}}$$

$$Bh = \mu_0 nih$$

$$\boxed{B = \mu_0 ni}$$

$$\xrightarrow{\text{magnetic field}}_{\text{inside a solenoid}} (31)$$

The uniform magnetic field inside a long solenoid is proportional to the current i in the solenoid, and the number of turns per unit length, n.

inside a solenoid

## **Exercise 7**

We will so often be using solenoids later in the course, that you should be able to derive the formula  $B = \mu_0 ni$ , starting from Ampere's law without looking at notes. This is a good time to practice. Take a blank sheet of paper, sketch a solenoid of length L with N turns. Then close the text and any notes, and derive the formula for B.

We have mentioned that equations like  $\oint \vec{B}_1 \cdot d\vec{\ell} = \mu_0 i_{enclosed}$  are meaningless hen scratching until you know how to use them. The best way to do that is learn worked examples along with the equation. Two good examples for Ampere's law are to be able to calculate the magnetic field inside a wire (Exercise 4), and to be able to derive the magnetic field inside a solenoid. If you can do these two derivations without looking at notes, you should have a fairly good grasp of the law.



### Figure 26

If we go around the wrong way, we just get two minus signs and all the results are the same. Here we went around the path so that our thumb pointed opposite to the direction of the current through the path. As a result the magnetic field in the solenoid points opposite to the direction of the path in the solenoid.

## **One More Right Hand Rule**

If we really want to be careful about minus signs (and it is not always necessary), we have to say how the sign of  $i_{enclosed}$  is evaluated in Figure (25). If, as in Figure (26) we reversed the direction of our path, then on side (1)  $\vec{B} \cdot d\vec{\ell}$  is negative because our path is going in the opposite direction to  $\vec{B}$ . Thus for this path the complete integral  $\oint \vec{B} \cdot d\vec{\ell}$  is negative, and somehow our  $i_{enclosed}$ must also be negative, so that we get the same answer we got for Figure (25).

If we curl the fingers of our *right hand* in the direction that we go around the path, then in Figure (25) our thumb points up parallel to the current through the path, and in Figure (26) our thumb points down, opposite to the current. If we define the direction indicated by our right hand thumb as the positive direction *through* the path, as shown in Figure (27), then the current is going in a positive direction in Figure (25) but in a negative direction in Figure (26). This gives us a negative  $i_{enclosed}$  for Figure (26) which goes along with the minus sign we got in the evaluation of  $\oint \vec{B} \cdot d\vec{l}$ .

By now you should be getting the idea of how we define directions in magnetic formula. *Always use your right hand*. After a while you get so used to using your right hand that you do not have to remember the individual right hand rules.



### Figure 27

In general, we use the right hand convention to associate a positive direction around a path to a positive direction through a path.

# The Toroid

If we take a long solenoid, bend it in a circle and fit the ends together, we get what is called a toroid shown in Figure (28). The great advantage of a toroid is that there are no end effects. In the straight solenoid the magnetic field at the ends fanned out into space as seen in our iron filing map of Figure (28-23). With the toroid there are no ends. The field is completely confined to the region inside the toroid and there is essentially no field outside. For this reason a toroid is an ideal magnetic field storage device.

It is easy to use Ampere's law to calculate the magnetic field inside the toroid. In Figure (28) we have drawn a path of radius r inside the toroid. Going around this path in the same direction as B, we immediately get

$$\oint \vec{B} \cdot d\vec{\ell} = B \times 2\pi r \tag{32}$$

because B is constant in magnitude and parallel to  $d\ell$ .



$$i_{\text{enclosed}} = \text{Ni}$$
 (33)

Using Equations (32) and (33) in Ampere's law gives

$$\oint \vec{B} \cdot d\vec{\ell} = \mu_0 i_{\text{enclosed}}$$

$$B \times 2\pi r = \mu_0 \text{Ni}$$

$$B = \frac{\mu_0 \text{Ni}}{2\pi r}$$

$$magnetic field of a toroid (34)$$

Note that  $N/2\pi r$  is the number of turns per unit length, n, so that Equation (34) can be written  $B = \mu_0 ni$ which is the solenoid formula of Equation (31). To a good approximation the field in a toroid is the same as in the center of a straight solenoid.

The **derivation** of Equation (34) is so easy and such a good illustration of the use of Ampere's law that it should be remembered as an example of Ampere's law.



### Figure 28

When the solenoid is bent into the shape of a toroid, there are no end effects. The magnetic field is confined to the region inside the toroid, and Ampere's law is easily applied. (You should remember this as an example of the use of Ampere's law.)

## **Exercise 8**

Figure 29 shows a 400-turn solenoid that is 47.5 cm long and has a diameter of 2.54 cm. (The 10 turns of wire wrapped around the center are for a later experiment.) Calculate the magnitude of the magnetic field B near the center of the solenoid when the wire carries a current of 3 amperes. (Give your answer in tesla and gauss.)

## **Exercise 9**

Figure 30 shows the toroidal solenoid that we use in several experiments later on. The coil has 696 turns wound on a 2.6 cm diameter plastic rod bent into a circle of radius 21.5 cm. What is the strength of the magnetic field inside the coil when a current of 1 amp is flowing through the wire? (Give your answer in tesla and gauss.)



Figure 29

A 400 turn straight solenoid 47.5 cm long, wound on a 2.54 cm diameter rod.



**Figure 30** A 696 turn toroidal solenoid wound on a 2.6 cm diameter plastic rod bent into a circle of radius 21.5 cm.

# Chapter 30 Faraday's Law

In this chapter we will discuss one of the more remarkable, and in terms of practical impact, important laws of physics – Faraday's law. This law explains the operation of the air cart speed detector we have used in air track experiments, the operation of AC voltage generators that supply most of the electrical power in the world, and transformers and inductors which are important components in the electronic circuits in radio and television sets.

In one form, Faraday's law deals with the line integral  $\oint \vec{E} \cdot d\vec{\ell}$  of an electric field around a closed path. As an introduction we will begin with a discussion of this line integral for electric fields produced by static charges. (Nothing very interesting happens there.) Then we will analyze an experiment that is similar to our air cart speed detector to see why we get a voltage proportional to the speed of the air cart. Applying the principle of relativity to our speed detector, i.e., riding along with the air cart gives us an entirely new picture of the behavior of electric fields, a behavior that is best expressed in terms of the line integral  $\oint \vec{E} \cdot d\vec{\ell}$ . After a discussion of this behavior, we will go through some practical applications of Faraday's law.

# ELECTRIC FIELD OF STATIC CHARGES

In this somewhat formal section, we show that  $\oint \vec{E} \cdot d\vec{\ell} = 0$  for the electric field of static charges. With this as a background, we are in a better position to appreciate an experiment in which  $\oint \vec{E} \cdot d\vec{\ell}$  is not zero.

In Figure (1), we have sketched a closed path through the electric field  $\vec{E}$  of a point charge, and wish to calculate the line integral  $\oint \vec{E} \cdot d\vec{\ell}$  for this path. To simplify the calculation, we have made the path out of arc and radial sections. But as in our discussion of Figure 29-13, we can get arbitrarily close to any path using arc and radial sections, thus what we learn from the path of Figure (1) should apply to a general path.

Because the electric field is radial,  $\vec{E}$  is perpendicular to  $d\vec{\ell}$  and  $\vec{E} \cdot d\vec{\ell}$  is zero on the arc sections. On the radial sections, for every step out where  $\vec{E} \cdot d\vec{r}$  is positive there is an exactly corresponding step back where  $\vec{E} \cdot d\vec{r}$ is negative. Because we come back to the starting point, we take the same steps back as we took out, all the radial  $\vec{E} \cdot d\vec{r}$  cancel and we are left with  $\oint \vec{E} \cdot d\vec{\ell} = 0$ for the electric field of a point charge.



### Figure 1

Closed path through the electric field of a point charge. The product  $\vec{E} \cdot d\vec{\ell}$  is zero on the arc sections, and the path goes out just as much as it comes in on the radial sections. As a result  $\oint \vec{E} \cdot d\vec{\ell} = 0$  when we integrate around the entire path.

Now consider the distribution of fixed point charges shown in Figure (2). Let  $\vec{E}_1$  be the field of  $Q_1$ ,  $\vec{E}_2$  of  $Q_2$ , etc. Because an electric field is the force on a unit test charge, and because forces add as vectors, the total electric field  $\vec{E}$  at any point is the vector sum of the individual fields at that point

$$\vec{E} = \vec{E}_1 + \vec{E}_2 + \vec{E}_3 + \vec{E}_4 + \vec{E}_5$$
 (1)

We can now use Equation (1) to calculate  $\oint \vec{E} \cdot d\vec{\ell}$ around the closed path in Figure (2). The result is

$$\oint \vec{\mathbf{E}} \cdot d\vec{\ell} = \oint \left( \vec{\mathbf{E}}_1 + \vec{\mathbf{E}}_2 + \dots + \vec{\mathbf{E}}_5 \right) \cdot d\vec{\ell}$$
$$= \oint \vec{\mathbf{E}}_1 \cdot d\vec{\ell} + \dots + \oint \vec{\mathbf{E}}_5 \cdot d\vec{\ell}$$
(2)

But  $\oint \vec{E}_1 \cdot d\vec{\ell} = 0$  since  $\vec{E}_1$  is the field of a point charge, and the same is true for  $\vec{E}_2 \dots \vec{E}_5$ . Thus the right side of Equation (2) is zero and we have

$$\oint \vec{E} \cdot d\vec{\ell} = 0 \quad \begin{cases} \text{for the field } \vec{E} \text{ of} \\ \text{any distribution of} \\ \text{static charges} \end{cases} (3)$$

Equation (3) applies to any distribution of static charges, a point charge, a line charge, and static charges on conductors and in capacitors.





Closed path in a region of a distribution of point charge. Since  $\oint \vec{E} \cdot d\vec{\ell} = 0$  is zero for the field of each point charge alone, it must also be zero for the total field  $\vec{E} = \vec{E}_1 + \vec{E}_2 + \vec{E}_3 + \vec{E}_4 + \vec{E}_5$ 

# A MAGNETIC FORCE EXPERIMENT

Figures (3a,b) are two views of an experiment designed to test for the magnetic force on the conduction electrons in a moving copper wire. We have a wire loop with a gap and the loop is being pulled out of a magnet. At this instant only the end of the loop, the end opposite the gap, is in the magnetic field. It will soon leave the field since it is being pulled out at a velocity  $\vec{v}$  as shown.

In our earlier discussions we saw that a copper atom has two loosely bound conduction electrons that are free to flow from one atom to another in a copper wire. These conduction electrons form a negatively charged electric fluid that flows in a wire much like water in a pipe.

Because of the gap we inserted in the wire loop of Figure (3), the conduction electrons in this loop cannot flow. If we move the loop, the conduction electrons must move with the wire. That means that the conduction electrons have a velocity  $\vec{v}$  to the right as shown, perpendicular to the magnetic field which is directed into the page. Thus we expect that there should be a magnetic force

$$\vec{F}_{mag} = -\vec{ev} \times \vec{B}$$
 (4)

acting on the electrons. This force will be directed down as shown in Figure (3b).

Since the gap in the loop does not allow the conduction electrons to flow along the wire, how are we going to detect the magnetic force on them? There is no net force on the wire because the magnetic field exerts an equal and opposite force on the positive copper ions in the wire.

Our conjecture is that this magnetic force on the conduction electrons would act much like the gravitational force on the water molecules in a static column of water. The pressure at the bottom of the column is higher than the pressure at the top due to the gravitational force. Perhaps the pressure of the negatively



**Figure 3a** Wire loop moving through magnetic field of iron magnet.



### **Figure 3b**

When you pull a wire loop through a magnetic field, the electrons, moving at a velocity  $\vec{v}$  with the wire, feel a magnetic force  $\vec{F}_B = (-e)\vec{v} \times \vec{B}$  if they are in the field. This force raises the pressure of the electron fluid on the bottom of the loop and reduces it on the top, creating a voltage V across the gap. The arrow next to the voltmeter indicates a voltage rise for positive charge, which is a voltage drop for negative charge.

charged electric fluid is higher at the bottom of the loop than the top due to the magnetic force.

To find out if this is true, we use an electrical pressure gauge, which is a voltmeter. A correctly designed voltmeter measures an electrical pressure drop without allowing any current to flow. Thus we can place the voltmeter across the gap and still not let the conduction electrons flow in the loop.

If our conjecture is right, we should see a voltage reading while the magnetic force is acting. Explicitly there should be a voltage reading while the wire is moving and one end of the loop is in the magnetic field as shown. The voltage should go to zero as soon as the wire leaves the magnetic field. If we reverse the direction of motion of the loop, the velocity  $\vec{v}$  of the conduction electrons is reversed, the magnetic force -  $e\vec{v} \times \vec{B}$  should also be reversed, and thus the sign of the voltage on the voltmeter should reverse. If we oscillate the wire back and forth, keeping one end in the magnetic field, we should get an oscillating voltage reading on the meter.

The wonderful thing about this experiment is that all these predictions work precisely as described. There are further simple tests like moving the loop faster to get a stronger magnetic force and therefore a bigger voltage reading. Or stopping the wire in the middle of the magnetic field and getting no voltage reading. They all work! The next step is to calculate the magnitude of the voltage reading we expect to see. As you follow this calculation, do not worry about the sign of the voltage V because many sign conventions (right hand rules, positive charge, etc.) are involved. Instead concentrate on the basic physical ideas. (In the laboratory, the sign of the voltage V you read on a voltmeter depends on how you attached the leads of the voltmeter to the apparatus. If you wish to change the sign of the voltage reading, you can reverse the leads.)

Since voltage has the dimensions of the potential energy of a unit test charge, the magnitude of the voltage in Figure (3) should be the strength of the force on a unit test charge,  $(-e) \vec{v} \times \vec{B}$  with (-e) replaced by 1, times the height h over which the force acts. This height h is the height of the magnetic field region in Figure (3). Since v and B are perpendicular,  $|\vec{v} \times \vec{B}| = vB$ and we expect the voltage V to be given by

$$V = \begin{pmatrix} \text{force on unit} \\ \text{test charge} \end{pmatrix} \times \begin{pmatrix} \text{distance over} \\ \text{which force acts} \end{pmatrix}$$
$$V = vB \times h \qquad \qquad \begin{array}{c} \text{voltage V on loop} \\ \text{moving at speed} \\ \text{v through field B} \end{array}$$
(5)

**Figure 3c** *Pulling the coil out of the magnet* 



# **AIR CART SPEED DETECTOR**

The air cart velocity detector we have previously discussed, provides a direct verification of Equation (5). The only significant difference between the air cart speed detector and the loop in Figure (3) is that the speed detector coil has a number of turns (usually 10). In order to see the effect of having more than one turn in the coil, we show a two turn coil being pulled out of a magnetic field in Figure (4).

Figure (4) is beginning to look like a plumbing diagram for a house. To analyze the diagram, let us start at Position (1) at the top of the voltmeter and follow the wire all the way around until we get to Position (6) at the bottom end of the voltmeter. When we get to Position (2), we enter a region from (2) to (3) where the magnetic force is increasing the electron fluid pressure by an amount vBh, as in Figure (3).

Now instead of going directly to the voltmeter as in Figure (3), we go around until we get to Position (4)

Ē

E

Х

(2)

хх

ХХ

ХХ

ХХ

X (3) Ē X



(1)

(6)

Ì

volt

meter

 $\overrightarrow{\mathsf{v}}$ 

where we enter another region, from (4) to (5), where the magnetic force is increasing the fluid pressure. We get another increase of vBh, and then go to Position (6) at the bottom of the voltmeter. In Figure (4) we have two voltage rises as we go around the two loops, and we should get twice the reading on the voltmeter.

V = 2vBh	voltage reading
	for 2 loops

It is an easy abstraction to see that if our coil had N turns, the voltage rise would be N times as great, or

	voltage on an N turn	
V = NvBh	coil being pulled out	(6)
	of a magnetic field	

Adding more turns is an easy way to increase or



A two turn loop being pulled through a magnetic field. With two turns we have twice as much force pushing the electric fluid toward the bottom of the gap giving twice the voltage V.

(5)

The setup for the air cart speed detector is shown in Figure (6). A multi turn coil, etched on a circuit board as shown in Figure (5), is mounted as a sail on top of an air cart. Suspended over the air cart are two angle iron bars with magnets set across the top as shown. This produces a reasonably uniform magnetic field that goes across from one bar to the other as seen in the end view of Figure (6).

In Figure (7), we show the experiment of letting the cart travel at constant speed through the velocity detector. In the initial position (a), the coil has not yet reached the magnetic field and the voltage on the coil is zero, as indicated in the voltage curve at the bottom of the figure.



Figure 5

The multi turn coil that rides on the air cart. (Only 5 turns are shown.)



The situation most closely corresponding to Figure (4) is position (d) where the coil is leaving the magnet. According to Equation (6), the voltage at this point should be given by V = NvBh, where N = 10 for our 10 turn coil, v is the speed of the carts, B is the strength of the magnetic field between the angle iron bars, and h is the average height of the coils. (Since the coils are drawn on a circuit board the outer loop has the greatest height h and the inner loop the least.) The first time you use this apparatus, you can directly measure V, N, v and h and use Equation (6) to determine the magnetic field strength B. After that, you know the constants N, B and h, and Equation (6) written as

$$\mathbf{v} = \mathbf{V} \times \left(\frac{1}{\mathbf{NBh}}\right) \tag{6a}$$

gives you the cart's speed in terms of the measured voltage V. Equation (6a) explains why the apparatus acts as a speed detector.

Let us look at the voltage readings for the other cart positions. The zero readings at Positions (a) and (e) are easily understood. None of the coil is in the magnetic field and therefore there is no magnetic force or voltage.



## Figure 6

The Faraday velocity detector. The apparatus is reasonably easy to build. We first constructed a 10 turn coil by etching the turns of the coil on a circuit board. This was much better than winding a coil, for a wound coil tends to have wrinkles that produce bumps in the data. Light electrical leads, not shown, go directly from the coil to the oscilloscope. The coil is mounted on top of an air cart and moves through a magnetic field produced by two pieces of angle iron with magnets on top as shown. Essentially we have reproduced the setup shown in Figures 3 and 4, but with the coil mounted on an air cart. As long as the coil remains with one end in the magnetic field and the other outside, as shown in (b), there will be a voltage on the leads to the coil that is proportional to the velocity of the cart.

### Figure 6c

Velocity detector apparatus. The magnetic field goes across, between the two pieces of angle iron. The coil, mounted on a circuit board, is entering the magnetic field.



## Figure 7

Voltage on the coil as it moves at constant speed through the magnetic field. At position (a) the coil has not yet reached the field and there is no voltage. At position (b) one end of the cart is in the field, the other outside, and we get a voltage proportional to the speed of the cart. At (c) there is no voltage because both ends of the cart are in the magnetic field and the magnetic force on the two ends cancel. (There is no change of magnetic flux at this point.) At (d), the other end alone is inside the field, and we get the opposite voltage from the one we had at (b). (Due to the thickness of the coil and fringing of the magnetic field, the voltage rises and falls will be somewhat rounded.)



We need a closer look to understand the changes in voltage, when all or part of the coil is inside the magnetic field. This situation, for a one turn coil, is illustrated in Figure (8). For easier interpretation we have moved the gap and voltmeter to the bottom of the coil as shown. It turns out that it does not matter where the gap is located, we get the same voltage reading. We have also labeled the figures (b), (c), and (d) to correspond to the positions of the air cart in Figure (7).

In Figure (8c) where both ends of the coil are in the magnetic field, the conduction electrons are being pulled down in both ends and the fluid is balanced. The electron fluid would not flow in either direction if the gap were closed, thus there is no pressure across the gap and no voltage reading. In contrast, in Figure (8d) where only the left end of the coil is in the magnetic field, the magnetic force on the left side would cause the conduction electrons to flow counterclockwise around the loop if it were not for the gap. There must be an electric pressure or voltage drop across the gap to prevent the counterclockwise flow. This voltage drop is what we measure by the voltmeter.

In Figure (8b), where the coil is entering the magnetic field, the magnetic force on the right side of the coil would try to cause a clockwise flow of the conduction electrons. We should get a pressure or voltage opposite to Figure (8d) where the coil is leaving. This reversal in voltage is seen in the air cart experiment of Figure (7), as the cart travels from (b) to (d).

Note that in Figure (8), where the horizontal sections of the coil are also in the magnetic field, the magnetic force is across rather than along the wire in these sections. This is like the gravitational force on the fluid in a horizontal section of pipe. It does not produce any pressure drops.



(b) coil entering magnetic field



(c) coil completely in magnetic field



(d) coil leaving magnetic field

## Figure 8

When the coil is completely in the magnetic field, the magnetic force on the electrons in the left hand leg (1) is balanced by the force on the electrons in the right hand leg (2), and there is no net pressure or voltage across the gap. When the coil is part way out, there is a voltage across the gap which balances the magnetic force on the electrons. The sign of the voltage depends upon which leg is in the magnetic field.

## A RELATIVITY EXPERIMENT

Now that we have seen, from Figure (7), extensive experimental evidence for the magnetic force on the conduction electrons in a wire, let us go back to Figure (3) where we first considered these forces, and slightly modify the experiment. Instead of pulling the coil out of the magnet, let us pull the magnet away from the coil as shown in Figure (9b).

In Figure (9a) we have redrawn Figure (3), and added a stick figure to represent a student who happens to be walking by the apparatus at the same speed that we are pulling the coil out of the magnet. To this moving observer, the coil is at rest and she sees the magnet moving to the left as shown in (9b). In other words, pulling the magnet away from the coil is precisely the same experiment as pulling the coil from the magnet, except it is viewed by a moving observer.

The problem that the moving observer faces in Figure (9b) is that, to her, the electrons in the coil are at rest. For her the electron speed is v = 0 and the magnetic force  $\vec{F}_B$ , given by

$$\vec{F}_{B} = (-e)\vec{v} \times \vec{B} = 0$$
 (for Figure 9b) (7)

is zero! Without a magnetic force to create the pressure in the electrical fluid in the wire, she might predict that there would be no voltage reading in the voltmeter.

But there is a voltage reading on the voltmeter! We have used this voltage to build our air cart velocity detector. If the voltmeter had a digital readout, for example, then it is clear that everyone would read the same number no matter how they were moving, whether they were like us moving with the magnet (9a), or like her moving with the coil (9b). In other words, she has to find some way to explain the voltage reading that she must see. The answer she needs lies in the Lorentz force law that we discussed in Chapter 28. This law tells us the total electromagnetic force on a charge q due to either electric or magnetic fields, or both. We wrote the law in the form

$$\vec{F} = q\vec{E} + q\vec{v} \times \vec{B}$$
(28-20)

where  $\vec{E}$  and  $\vec{B}$  are the electric and magnetic fields acting on the charge.



### Figure 9

The only difference between (a) and (b) is the point of view of the observer. In (a) we see a magnetic force  $\vec{F}_B = (-e) \vec{v} \times \vec{B}$  because the electrons are moving at a speed v through a magnetic field  $\vec{B}$ . To the observer in (b), the magnet is moving, not the electrons. Since the electrons are at rest, there is no magnetic force on them. Yet the voltmeter reading is the same from both points of view. Let us propose that the Lorentz force law is generally correct even if we change coordinate systems. In Figure (9a) where we explained everything in terms of a magnetic force on the conduction electrons, there was apparently no electric field and the Lorentz force law gave

$$\vec{F} = q\vec{E} + q\vec{v} \times \vec{B}$$
$$= (-e)\vec{v} \times \vec{B} \qquad \begin{pmatrix} in \ Figure \ 9a, \\ \vec{E} = 0 \end{pmatrix}$$
(8a)

In Figure (9b), where  $\vec{v} = 0$ , we have

$$\vec{F} = q\vec{E} + q\vec{v} \times \vec{B}$$
$$= (-e)\vec{E} \qquad \begin{pmatrix} in \ Figure \ 9b, \\ \vec{v} = 0 \end{pmatrix}$$
(8b)

In other words, we will assume that the magnetic force of Figure (9a) has become an electric force in Figure (9b). Equating the two forces gives

$$\vec{E} \begin{pmatrix} That \ should \ be \\ in \ Figure \ 9b \end{pmatrix} = \vec{v} \times \vec{B} \begin{pmatrix} From \\ Figure \ 9a \end{pmatrix}$$
(9)

In Figure (9c) we have redrawn Figure (9b) showing an electric field causing the force on the electrons. Because the electrons have a negative charge, the electric field must point up in order to cause a downward force.

That the magnetic force of Figure (9a) becomes an electric force in Figure (9c) should not be a completely surprising result. In our derivation of the magnetic force law, we also saw that an electric force from one point of view was a magnetic force from another point of view. The Lorentz force law, which includes both electric and magnetic forces, has the great advantage that it gives the correct electromagnetic force from any point of view.

## Exercise 1

Equation (9) equates  $\vec{E}$  in Figure (9c) with  $\vec{v} \times \vec{B}$  in Figure (9a). Show that  $\vec{E}$  and  $\vec{v} \times \vec{B}$  point in the same direction.



### Figure 9c

From the point of view that the coil is at rest, the downward force on the electrons in the coil must be produced by an upward directed electric field.

## FARADAY'S LAW

An experiment whose results may be surprising, is shown in Figure (10). Here we have a magnetic field produced by an electromagnet so that we can turn  $\vec{B}$  on and off. We have a wire loop that is large enough to surround but not lie in the magnetic field, so that  $\vec{B} = 0$ all along the wire. Again we have a gap and a voltmeter to measure any forces that might be exerted on the conduction electrons in the wire.

We have seen that if we pull the wire out of the magnet, Figure (9a), we will get a voltage reading while the loop is leaving the magnetic field. We have also seen, Figure (9c), that we get a voltage reading if the magnetic field is pulled out of the loop. In both cases we started with a magnetic field through the loop, ended up with no magnetic field through the loop, and got a reading on the voltmeter while the amount of magnetic field through the loop was decreasing.

Now what we are going to do in Figure (10) is simply shut off the electromagnet. Initially we have a magnetic field through the loop, finally no field through the loop. It may or may not be a surprise, but *when we shut off the magnetic field, we also get a voltage reading*. We get a voltage reading if we pull the loop out of the field, the field out of the loop, or shut off the field. We are seeing that we *get a voltage reading whenever we change the amount of magnetic field, the flux of magnetic field, through the loop*.



### Figure 10

Here we have a large coil that lies completely outside the magnetic field. Thus there is no magnetic force on any of the electrons in the coil wire. Yet when we turn the magnet on or off, we get a reading in the volt meter.

## 30-11

### Magnetic Flux

In our discussion of velocity fields and electric fields, we used the concept of the flux of a field. For the velocity field, the flux  $\Phi_v$  of water was the volume of water flowing per second past some perpendicular area  $A_{\perp}$ . For a uniform stream moving at a speed v, the flux was  $\Phi_v = vA_{\perp}$ . For the electric field, the formula for flux was  $\Phi_E = EA_{\perp}$ .

In Figures (9 and 10), we have a magnetic field that "flows" through a wire loop. Following the same convention that we used for velocity and electric fields, we will define the magnetic flux  $\Phi_B$  as the strength of the field  $\vec{B}$  times the perpendicular area  $A_{\perp}$  through which the field is flowing

$$\Phi_{\rm B} = {\rm BA}_{\perp} \qquad \begin{array}{c} {\rm Definition \ of} \\ {\rm magnetic \ flux} \end{array} \tag{10}$$

In both figures (9) and (10), the flux  $\Phi_B$  through the wire loop is decreasing. In Figure (9),  $\Phi_B$  decreases because the perpendicular area  $A_{\perp}$  is decreasing as the loop and the magnet move apart. In Figure (10), the flux  $\Phi_B$  is decreasing because  $\vec{B}$  is being shut off. The important observation is that whenever the flux  $\Phi_B$  through the loop decreases, whatever the reason for the change may be, we get a voltage reading V on the voltmeter.

# **One Form of Faraday's Law**

The precise relationship between the *voltage* and the *change in the magnetic flux* through the loop is found from our analysis of Figure (9) where the loop and the magnet were pulled apart. We got a voltage given by Equation (5) as

$$V = vBh$$
(5)

Let us apply Equation (5) to the case where the magnet is being pulled out of the loop as shown in Figure (11). In a time dt, the magnet moves to the left a distance dx given by

$$dx = vdt (11)$$

and the area of magnetic field that has left the loop, shown by the cross hatched band in Figure (11), is

$$dA = hdx = \begin{cases} \text{area of magnetic} \\ \text{field that has} \\ \text{left the loop} \end{cases}$$
(12)

This decrease in area causes a decrease in the magnetic flux  $\Phi_B = BA_{\perp}$  through the loop. The change in flux  $d\Phi_B$  is given by



## Figure 11

As the magnet and the coil move away from each other, the amount of magnetic flux through the coil decreases. When the magnet has moved a distance dx, the decrease in area is hdx, and the magnetic flux decreases by  $B \times hdx$ .

$$d\Phi_{\rm B} = -BdA = -Bhdx$$
$$= -Bhvdt$$
(13)

where the - sign indicates a reduction in flux, and we used Equation (11) to replace dx by vdt.

Dividing both sides of Equation (13) by dt gives

$$\frac{\mathrm{d}\Phi_{\mathrm{B}}}{\mathrm{d}t} = -\mathrm{B}\mathrm{h}\mathrm{v} \tag{14}$$

But Bhv is just our voltmeter reading. Thus we get the surprisingly simple formula

$$V = -\frac{d\Phi_{\rm B}}{dt} \qquad One form of Faraday's law \qquad (15)$$

Equation (15) is one form of Faraday's law.

Equation (15) has a generality that goes beyond our original analysis of the magnetic force on the conduction electrons. It makes no statement about what causes the magnetic flux to change. We can pull the loop out of the field as in Figure (9a), the field out of the loop as in Figure (9b), or shut the field off as in Figure (10). In all three cases Equation (15) predicts that we should see a voltage, and we do.

If we have a coil with more than one turn, as we had back in Figure (4), and put a voltmeter across the ends of the coil, then we get N times the voltage, and Equation (15) becomes

$$V = N \left( -\frac{d\Phi_{B}}{dt} \right) \qquad for a coil \\ with N turns \qquad (15a)$$

provided  $d\Phi_B/dt$  is the rate of change of magnetic flux in each loop of the coil.

### Exercise 2

Go back to Figure (7) and explain the voltage plot in terms of the *rate of change of the flux of magnetic field* through the coil riding on top of the air cart.

# **A Circular Electric Field**

In Figure (10), where we shut the magnet off and got a voltage reading on the voltmeter, there must have been some force on the electrons in the wire to produce the voltage. Since there was no magnetic field out at the wire, the force must have been produced by an electric field. We already have a hint of what that electric field looks like from Figure (9c). In that figure, we saw that the moving magnetic field acting on the electrons on the left side of the wire loop.

To figure out the shape of the electric field produced when we shut off the magnet, consider Figure (12), where we have a circular magnet and a circular loop of wire . We chose this geometry so that the problem would have circular symmetry (except at the gap in the loop). To produce the same kind of voltage V that we have seen in the previous experiments, the electric field at the wire must be directed up on the left hand side, as it was in Figure (9c). But because of the circular symmetry of the setup in Figure (12), the upwardly directed electric field on the left side, which is parallel to the wire, must remain parallel to the wire as we go around the wire loop. In other words, the only way we can have an upwardly directed electric field acting on the electrons on the left side of the loop, and maintain circular symmetry, is to have the electric field go in a circle all the way around the loop as shown in Figure (12).

We can determine the strength of this circular electric field, by figuring out how strong an electric field must act on the electrons in the wire, in order to produce the voltage V across the gap. We then use Equation (15) to relate this voltage to the rate of change of the magnetic flux through the loop.



## Figure 12

When the magnetic field in the magnet is turned off, a circular electric field is generated. This electric field exerts a force on the electrons in the wire, creating a pressure in the electric fluid that is recorded as a voltage pulse by the voltmeter.

Recall that the definition of electric voltage used in deriving Equation (5) was

$$\mathbf{V} = \begin{pmatrix} \text{force on unit} \\ \text{test charge} \end{pmatrix} \times \begin{pmatrix} \text{distance over} \\ \text{which force acts} \end{pmatrix}$$

For Figure (12), the force on a unit test charge is the electric field  $\vec{E}$ , and this force acts over the full circumference  $2\pi r$  of the wire loop. Thus the voltage V across the gap is

$$V = E \times 2\pi r$$

Equating this voltage to the rate of change of magnetic flux through the wire loop gives

$$V = E \times 2\pi r = -\frac{d\Phi_B}{dt}$$
(16)

Equation (16) tells us that the faster the magnetic field dies, i.e. the greater  $d\Phi_B/dt$ , the stronger the electric field  $\vec{E}$  produced.

## Line Integral of $\vec{E}$ around a Closed Path

In Figure (13) we have removed the wire loop and volt meter from Figure (12) so that we can focus our attention on the circular electric field produced by the decreasing magnetic flux. This is not the first time we have encountered a circular field. The velocity field of a vortex and the magnetic field of a straight current carrying wire are both circular. We have redrawn Figure (29-10) from the last chapter, showing the circular magnetic field around a wire.

The formula for the strength of the magnetic field in Figure (29-10) is

$$B \times 2\pi r = \mu_0 i \tag{28-18}$$

a result we derived back in Equation 28-18. This should be compared with the formula for the strength of the electric field in Figure (13)

$$\mathbf{E} \times 2\pi \mathbf{r} = -\frac{\mathrm{d}\Phi_{\mathrm{B}}}{\mathrm{d}t} \tag{16}$$



Circular electric field around a changing magnetic flux.

**Figure 29-10** *Circular magnetic field around an electric current.*  In our discussion of Ampere's law, we called  $\mu_0$  i the "source" of the circular magnetic field. By analogy, we should think of the rate of change of magnetic flux,  $- d\Phi_B/dt$ , as the "source" of the circular electric field.

In Chapter 29, we generalized Ampere's law by replacing  $B * 2\pi r$  by the line integral  $\oint \vec{B} \cdot d\vec{l}$  along a closed path around the wire. The result was

$$\oint \vec{B} \cdot d\vec{\ell} = \mu_0 i \qquad Amperes \ law \\ for \\ magnetic \ fields \qquad (29-18)$$

where the line integral can be carried out along any closed path surrounding the wire. Because of close analogy between the structure and magnitude of the magnetic field in Figure (29-10) and the electric field in Figure (13), we expect that the more general formula for the electric field produced by a changing magnetic flux is

$$\oint \vec{E} \cdot d\vec{\ell} = -\frac{d\Phi_B}{dt} \begin{cases} Faraday's law \\ for \\ electric fields \end{cases}$$
(17)

Equation 17 is the most general form of Faraday's law. It says that the line integral of the electric field around any closed path is equal to (minus) the rate of change of magnetic flux through the path.



**Figure 13a** Using Faraday's law to calculate  $\vec{E}$ .

# **USING FARADAY'S LAW**

Up until now we have been looking for arguments leading up to Faraday's law. Let us now reverse the procedure, treating Equation 17 as a basic law for electric fields, and see what the consequences are.

## **Electric Field of an Electromagnet**

As a beginning exercise in the use of Faraday's law, let us use Equation (17) to calculate the electric field of the electromagnet in Figure (13). We first argue that because of the circular symmetry, the electric field should travel in circles around the decreasing magnetic field. Thus we choose a circular path, shown in Figure (13a), along which we will calculate  $\oint \vec{E} \cdot d\vec{\ell}$ . Then using the assumption (because of circular symmetry) that  $\vec{E}$  is parallel to  $d\vec{\ell}$  and has a constant magnitude all the way around the circular path, we can write

$$\oint \vec{E} \cdot d\vec{\ell} = \oint E d\ell = E \oint d\ell = E 2\pi r \qquad (18)$$

Using this result in Equation (17) gives

$$\oint \vec{E} \cdot d\vec{\ell} = E \, 2\pi r = -\frac{d\Phi_B}{dt} \tag{19}$$

which is the result we had in Equation (16).

## **Right Hand Rule for Faraday's Law**

We can get the correct direction for  $\vec{E}$  with the following right hand rule. Point the thumb of your right hand in the direction of the magnetic field. If the magnetic flux is decreasing (if  $-d\Phi_B/dt$  is positive), then the fingers of your right hand curl in the direction of  $\vec{E}$ . If the magnetic flux is increasing, then  $\vec{E}$  points the other way. *Please practice this right hand rule on Figures* (13a), (9c), and (15).

# **Electric Field of Static Charges**

If all we have around are static electric charges, then there are no magnetic fields, no magnetic flux, and no changing magnetic flux. For this special case,  $d\Phi_B/dt = 0$  and Faraday's law gives

$$\oint \vec{E} \cdot d\vec{\ell} = 0 \quad \begin{cases} \text{for electric fields} \\ \text{produced by} \\ \text{static charges} \end{cases}$$
(20)

When the line integral of a force is zero around any closed path, we say that the force is *conservative*. (See Equation 29-12.) Thus we see that if we have only static electric charge (or constant magnetic fields), the electric field is a conservative field.

In contrast, if we have changing magnetic fields, if  $d\Phi_B/dt$  is not zero, the electric field is not conservative. This can lead to some rather interesting results which we will see in our discussion of a device called the *betatron*.



## Figure 14a

Cross-sectional view of a betatron, showing the central field  $\vec{B}_0$  and the field  $\vec{B}_r$  out at the evacuated doughnut. The relative strength of  $\vec{B}_0$  and  $\vec{B}_r$  can be adjusted by changing the shape of the electromagnet pole pieces.

# THE BETATRON

As we have mentioned before, when you encounter a new and strange equation like Faraday's law, it is essential to have an example that you know inside out that illustrates the equation. This transforms the equation from a collection of symbols into a set of instructions for solving problems and making predictions. One of the best examples to learn for the early form of Faraday's law, Equation (15a), was the air cart speed detector experiment shown in Figure (7). (You should have done Exercise 2 analyzing the experiment using Equation (15a).

The most direct example illustrating Faraday's law for electric fields, Equation (17), is the particle accelerator called the betatron. This device was used in the 1950s for study of elementary particles, and later for creating electron beams for medical research.

A cross-sectional view of the betatron is shown in Figure (14a). The device consists of a large electromagnet with a circular evacuated doughnut shaped chamber for the electrons. The circular shape of the electromagnet and the evacuated chamber are more clearly seen in the top view, Figure (14b). In that view we show the strong upward directed magnetic field  $\vec{B}_0$  in the gap and the weaker upward directed magnetic field  $\vec{B}_0$  in the gap and the evacuated doughnut.

The outer magnetic field  $B_r$  is required to keep the electrons moving along a circular orbit inside the evacuated chamber. This field exerts a force  $\vec{F}_B = (-e) \vec{v} \times \vec{B}_r$  that points toward the center of the circle and has a magnitude  $mv^2/r$  in order to produce the required radial acceleration. Thus  $B_r$  is given by

$$B_r = \frac{mv}{er}$$

which is our familiar formula for electrons moving along a circular path in a magnetic field. (*As a quick review, derive the above equation.*)

Since a magnetic field does no work we need some means of accelerating the electrons. In a synchrotron, shown in Figure (28-27), a cavity which produces an electric accelerating field is inserted into the electron's path. As an electron gains energy and momentum (mv) each time it goes through the cavity, the magnetic field B was increased so that the electron's orbital radius r = mv/eB remains constant. (The synchronizing of B with the momentum mv leads to the name synchrotron.)

In the betatron of Figure (14), we have a magnetic field  $B_r$  to keep the electrons in a circular orbit, and as the electrons are accelerated,  $B_r$  is increased to keep the electrons in an orbit of constant radius r. But what accelerates the electrons? There is no cavity as in a synchrotron.

Suppose that both  $B_0$  and  $B_r$  are increased simultaneously. In the design shown in Figure (14a),  $\vec{B}_0$  and  $\vec{B}_r$  are produced by the same electromagnet, so that we can increase both together by turning up the electromagnet. If the strong central field  $\vec{B}_0$  is increased, we have a large change in the magnetic flux through the electron orbit, and therefore by Faraday's law  $\oint \vec{E} \cdot d\vec{\ell} = -d\Phi_B/dt$  we must have a circular electric

field around the flux as shown in Figure (15), just as in Figure (13). This electric field is exactly parallel to the orbit of the electrons and accelerates them continuously as they go around.

What is elegant about the application of Faraday's law to the electrons in the betatron, is that  $\oint \vec{E} \cdot d\vec{\ell}$ , which has the dimensions of voltage, is the *voltage gained by an electron going once around the circular orbit. The energy gained is just this voltage in electron volts* 

energy gained  
(in eV) by electron  
going around once 
$$\left. = \oint \vec{E} \cdot d\vec{\ell} \right|$$
(21)

This voltage is then related to  $d\Phi_B/dt$  by Faraday's law.



### Figure 14b

Top view of the betatron showing the evacuated doughnut, the path of the electrons, and the magnetic fields  $\vec{B}_0$  in the center and  $\vec{B}_r$  out at the electron path. In order to keep the electrons moving on a circular path inside the doughnut, the magnetic force  $\vec{F}_B = (-e)\vec{v} \times \vec{B}_r$  must have a magnitude  $\vec{F}_B = mv^2/r$  where r is the radius of the evacuated doughnut.



### Figure 15

When the strong central field  $\vec{B}_0$  in the betatron is rapidly increased, it produces a circular electric field that is used to accelerate the electrons. The electric field  $\vec{E}$ is related to the flux  $\Phi_B$  of the central field  $\vec{B}_0$  by Faraday's law  $\oint \vec{E} \cdot d \, \ell = -d \, \Phi_B / dt$ . Let us consider an explicit example to get a feeling for the kind of numbers involved. In the 100 MeV betatron built by General Electric, the electron orbital radius is 84 cm, and the magnetic field  $B_0$  is cycled from 0 to .8 tesla in about 4 milliseconds. (The field  $B_0$  is then dropped back to 0 and a new batch of electrons are accelerated. The cycle is repeated 60 times a second.)

The maximum flux  $\Phi_{\rm m}$  through the orbit is

$$\Phi_{\rm m} = (B_0)_{\rm max} \pi r^2 = .8 \text{ tesla} \times \pi \times (.84\text{m})^2$$
$$\Phi_{\rm m} = 1.8 \text{ tesla m}^2$$

If this amount of flux is created in 4 milliseconds, then the average value of the rate of change of magnetic flux  $\Phi_B$  is

$$\frac{d\Phi_{\rm B}}{dt} = \frac{\Phi_{\rm m}}{.004 \text{ sec}} = \frac{1.8}{.004} = 450 \text{ volts}$$

Thus each electron gains 450 electron volts of kinetic energy each time it goes once around its orbit.

## Exercise 3

(a) How many times must the electron go around to reach its final voltage of 100 MeV advertised by the manufacturer?

(b) For a short while, until the electron's kinetic energy gets up to about the electron's rest energy  $m_0c^2$ , the electron is traveling at speeds noticeably less than c. After that the electron's speed remains very close to c. How many orbits does the electron have to make before its kinetic energy equals its rest energy? What fraction of the total is this?

(c) How long does it take the electron to go from the point that its kinetic energy equals its rest energy, up to the maximum of 100 MeV? Does this time fit within the 4 milliseconds that the magnetic flux is being increased?

# **TWO KINDS OF FIELDS**

At the beginning of the chapter we showed that the line integral  $\oint \vec{E} \cdot d\vec{\ell}$  around a closed path was zero for any electric field produced by static charges. Now we see that the line integral is not zero for the electric field produced by a changing magnetic flux. Instead it is given by Faraday's law  $\oint \vec{E} \cdot d\vec{\ell} = -d\Phi_B/dt$ . These results are shown schematically in Figure (16) where we are looking at the electric field of a charged rod in (16a) and a betatron in (16b).

In Figure (17), we have sketched a wire loop with a voltmeter, the arrangement we used in Figure (12) to measure the  $\oint \vec{E} \cdot d\vec{\ell}$ . We will call this device an " $\oint \vec{E} \cdot d\vec{\ell}$  meter". If you put the  $\oint \vec{E} \cdot d\vec{\ell}$  meter over the changing magnetic flux in Figure (16b), the voltmeter will show a reading of magnitude V =  $d\Phi_{\rm B}/dt$ . If we put the  $\oint \vec{E} \cdot d\vec{l}$  meter over the charged rod in Figure (16a), the meter reads V = 0. Thus we have a simple physical device, our  $\oint \vec{E} \cdot d\vec{\ell}$  meter, which can distinguish the radial field in Figure (16a) from the circular field in Figure (16b). In fact it can distinguish the circular field in (16b) from any electric field E whatsoever that we can construct from static charges. Our  $\oint \vec{E} \cdot d\vec{\ell}$  meter allows us to separate all electric fields into two kinds, those like the one in (16b) that can give a *non zero reading*, and those, produced by static charges, which give a *zero reading*.

Fields which register on our  $\oint \vec{E} \cdot d\vec{\ell}$  meter generally close on themselves like the circular fields in (16b). Since these fields do not appear to have sources, they are called sourceless or "*solenoidal*" fields. An  $\oint \vec{E} \cdot d\vec{\ell}$  meter is the kind of device we need to detect solenoidal fields.

The conservative fields produced by static charges never close on themselves. They always start on positive charge, end on negative charge, or come from or go to infinity. These fields diverge from point charges and thus are sometimes called *"divergent"* fields. Our  $\oint \vec{E} \cdot d\vec{\ell}$  meter does not work on the divergent fields because we always get a zero reading.





(a) Electric field of a static charge distribution has the property  $\oint \vec{E} \cdot d\vec{\ell} = 0$ 



(b) Electric field produced by a changing magnetic flux has  $\oint \vec{E} \cdot d\vec{\ell} = - d\vec{\Phi}_B/dt$ 

### Figure 16

Two kinds of electric field. Only the field produced by the changing magnetic flux has a non zero line integral.

Although the  $\oint \vec{E} \cdot d\vec{\ell}$  meter does not work on divergent fields, Gauss' law with the surface integral does. In a number of examples we used Gauss' law

$$\int_{\substack{\text{closed}\\\text{surface}}} \vec{E} \cdot d\vec{A} = \frac{Q_{\text{in}}}{\epsilon_0}$$
(29-5)

to calculate the electric field of static charges. We are seeing now that we *use a surface integral to measure divergent fields, and a line integral to measure solenoidal fields*. There are two kinds of electric fields, and we have two kinds of integrals to detect them.

It turns out to be a general mathematical theorem that any vector field can be separated into a purely divergent part and a purely solenoidal part. The field can be uniquely specified if we have both an equation involving a Gauss' law type surface integral to tell us the divergent part, and an equation involving a Faraday's law type line integral to tell us the solenoidal part.



### Figure 17

Wire loop and a volt meter can be used directly to measure  $\oint \vec{E} \cdot d\vec{\ell}$  around the loop. We like to call this apparatus an  $\oint \vec{E} \cdot d\vec{\ell}$  meter.
#### **Exercise 4**

a) Maxwell's equations are a set of equations that completely define the behavior of electric fields  $\vec{E}$  and magnetic fields  $\vec{B}$ . One of Maxwell's equations is Faraday's law

$$\oint \vec{E} \cdot d\vec{\ell} = -d\Phi_{B}/dt$$

which gives the line integral for the electric field. How many Maxwell equations are there? (How many equations will it take to completely define both  $\vec{E}$  and  $\vec{B}$ ?)

b) Are any of the other equations for electric and magnetic fields we have discussed earlier, candidates to be one of Maxwell's equations?

c) At least one of Maxwell's equations is missing – we have not discussed it. Can you guess what the equation is and write it down? Explain what you can about your guess.

d) Back in our early discussion of velocity fields and Gauss' law, we said that a *point source* for the velocity field of an incompressible fluid like water, was a small "magic" sphere in which water molecules were created. Suppose we do not believe in magic and assume that for real water there is no way that water molecules can be created or destroyed. Write down an integral equation for real water that expresses the fact that the  $\vec{v}_{real \ water}$  has no sources (that create water molecules) or sinks (that destroy them).

Do the best you can on these exercises now. Keep a record of your work, and see how well you did when we discuss the answers later in chapter 32.

## Note on our $\oint ec{m{E}} \cdot dec{\ell}$ meter

Back in Figure (17) we used a wire loop and a voltmeter as an  $\oint \vec{E} \cdot d\vec{\ell}$  meter. I.e., we are saying that the voltage reading V on the voltmeter gives us the integral of  $\vec{E}$  around the closed path defined by the wire loop. This is strictly true for a loop at rest, where the conduction electrons experience no magnetic force and all forces creating the electric pressure are caused by the electric field  $\vec{E}$ .

Earlier, in Figure (9), we had two views of an  $\oint \vec{E} \cdot d\vec{\ell}$ meter. In the bottom view, (9b) the loop is at rest and the voltage must be caused by an electric force. The moving magnetic field must have an electric field associated with it. But in Figure (9a) where the magnet is at rest, there is no electric field and the voltage reading is caused by the magnetic force on the conduction electrons in the moving wire. Strictly speaking, in Figure (9a) the wire loop and voltmeter are measuring a pressure caused by magnetic forces and not an  $\oint \vec{E} \cdot d\vec{\ell}$ . The wire loop must be at rest, the path for our line integral cannot move, if we are measuring  $\oint \vec{E} \cdot d\vec{\ell}$ .

In practice, however, it makes little difference whether we move the magnet or the loop, because the principle of relativity requires that we get the same voltage V.

## APPLICATIONS OF FARADAY'S LAW

The last few sections have been somewhat heavy on theory. To end this chapter on a more practical note, we will consider some simple applications of Faraday's law, one that has immense practical applications and another that we can use in the laboratory. First we will discuss the AC voltage generator which is used by most power stations throughout the world. We will also describe a field mapping experiment in which we use our  $\oint \vec{E} \cdot d\vec{\ell}$  meter to map the magnetic field of a pair of Helmholtz coils. In the next chapter Faraday's law is used to explain the operation of transformers and inductors that are common circuit elements in radio and television sets.



a) end view of a coil of wire rotating in a magnetic field



b) top view showing the coil of area A



c) Vector  $\vec{A}$  representing the area of the loop

#### Figure 18

An electric generator consists of a coil of wire rotating in a magnetic field.

## The AC Voltage Generator

In Figure (18) we have inserted a wire loop of area  $\vec{A}$  in the magnetic field  $\vec{B}$  of a magnet. We then rotate the coil at a frequency  $\omega$  about an axis of the coil as shown. We also attach a voltmeter to the coil, using sliding contacts so that the voltmeter leads do not twist as the coil spins.

As shown in Figure (19), as the loop turns, the magnetic flux changes sinusoidally from a maximum positive flux in (19a) to zero flux in (c) to a maximum negative flux in (d) to zero in (e). In (18c), we have shown the vector  $\vec{A}$  representing the area of the coil ( $\vec{A}$  points



#### Figure 19

The changing magnetic flux through the rotating loop. The general formula for  $\Phi_B$  is  $BA\cos\theta$  where  $\theta$  is the angle shown in (b), between the magnetic field and the normal to the loop. If the coil is rotating uniformly,

then  $\theta = \omega t$ , and  $\Phi_B = BA \cos \omega t$ 

perpendicular to the plane of the coil) and we can use our usual formula for magnetic flux to get

$$\Phi_{\rm B} = \vec{\rm B} \cdot \vec{\rm A} = {\rm B} {\rm A} \cos \theta \tag{22}$$

If the coil is rotating at a constant angular velocity  $\omega$ , then  $\theta = \omega t$  and we have

$$\Phi_{\rm B} = B \operatorname{Acos} \omega t \tag{23}$$

Differentiating Equation (23) with respect to time gives

$$\frac{d\Phi_{\rm B}}{dt} = -\omega BA\sin\omega t \tag{24}$$

Finally we use Faraday's law in the form

$$V = -\frac{d\Phi_B}{dt}$$
(15)

to predict that the voltage V on the voltmeter will be

$$V = \omega BA \sin \omega t \tag{25}$$

If we use a coil with N turns, we get a voltage N times as great, or

$$V = \omega NBA \sin \omega t = V_0 \sin \omega t$$
 (26)

where  $V_0$  is the amplitude of the sine wave as shown in Figure (20). Equation (26) shows that by rotating a coil in a magnetic field, we get an alternating or "AC" voltage. Power stations use this same principle to generate AC voltages.



**Figure 20** *Amplitude and period of a sine wave.* 

Equation 26 predicts that the voltage amplitude  $V_0$  produced by an N turn coil of area A rotating in a magnetic field B is

$$V_0 = \omega NBA \tag{27}$$

where the angular frequency  $\omega$  radians per second is related to the frequency f cycles per second and the period T seconds per cycle by

$$\omega \frac{\text{rad}}{\text{sec}} = 2\pi \frac{\text{rad}}{\text{cycle}} \times f \frac{\text{cycle}}{\text{sec}} = 2\pi \frac{\text{rad}}{\text{cycle}} \times \frac{1}{T \frac{\text{sec}}{\text{cycle}}}$$

#### **Exercise 5**

Suppose that you have a magnetic field B = 1 tesla, and you rotate the coil at 60 revolutions (cycles) per second. Design a generator that will produce a sine wave voltage whose amplitude is 120 volts.

#### **Exercise 6**

Figures (21a,b) show the voltage produced by a coil of wire rotating in a uniform magnetic field of a fairly large electromagnet. (The setup is similar to that shown in Figures 18 and 19.) The coil was square, 4 cm on a side, and had 10 turns. To go from the results shown in Figure (21a) to those shown in Figure (21b), we increased the rotational speed of the motor turning the coil. In both diagrams, we have selected one cycle of the output wave, and see that the frequency has increased from 10 cycles per second to nearly 31 cycles per second.

a) Explain why the amplitude of the voltage signal increased in going from Figure (21a) to (21b). Is the increase what you expected?

b) Calculate the strength of the magnetic field of the electromagnet used. Do you get the same answer using Figure (21a) and using Figure (21b)?



#### Figure 21

Voltage output from a coil rotating in a uniform magnetic field. The coil was 4 cm on a side, and had 10 turns. In each figure we have selected one cycle of the output wave, and see that the frequency of rotation increased from 10 cycles per second in a) to nearly 31 cycles per second in b).

#### Gaussmeter

Exercise 6 demonstrates one way to measure the strength of the magnetic field of a magnet. By spinning a coil in a magnetic field, we produce a voltage amplitude given by Equation 27 as  $V_0 = \omega NBA$ . Thus by measuring  $V_0$ ,  $\omega$ , N, and A, we can solve for the magnetic field B.

A device designed to measure magnetic fields is called a *gaussmeter*. A commercial gaussmeter, used in our plasma physics lab, had a small coil mounted in the tip of a metal tube as shown in Figure (22). A small motor also in the tube spun the coil at high speed, and the amplitude  $V_0$  of the coil voltage was displayed on a meter. The meter could have been calibrated using Equation (27), but more likely was calibrated by inserting the spinning coil into a known magnetic field.

In an attempt to measure the magnetic field in the Helmholtz coils used for our electron gun experiments, students have also built rotating coil gaussmeters. Despite excellent workmanship, the results were uniformly poor. The electrical noise generated by the sliding contacts and the motor swamped the desired signal except when B was strong. This approach turned out not to be the best way to measure  $\vec{B}$  in the Helmholtz coils.



#### Figure 22

A commercial gauss meter, which measures the strength of a magnetic field, has a motor and a rotating coil like that shown in Figure 18. The amplitude  $V_0$  of the voltage signal is displayed on a meter that is calibrated in gauss.

## **A Field Mapping Experiment**

To measure the magnetic field in the Helmholtz coils, it is far easier to "rotate the field" than the detector loop. That is, use an alternating current in the Helmholtz coils, and you will get an alternating magnetic field in the form

$$\mathbf{B} = \mathbf{B}_0 \sin \omega t \tag{28}$$

where w is the frequency of the AC current in the coils. Simply place a stationary detector loop in the magnetic field as shown in Figure (23) and the magnetic flux through the detector loop will be

$$\Phi_{\rm B} = \vec{B} \cdot \vec{A} = \vec{B}_0 \cdot \vec{A} \sin \omega t$$
 (29)

where A is the area of the detector loop. By Faraday's law, the voltage in the voltmeter or oscilloscope attached to the detector loop is given by

$$V = -\frac{d\Phi_{\rm B}}{dt} = -(\omega \vec{B}_0 \cdot \vec{A}) \cos \omega t$$
(30)

If our detector loop has N turns of wire, then the voltage will be N times as great, and the amplitude  $V_0$  we see on the oscilloscope screen will be

$$\mathbf{V}_0 = \mathbf{N}\boldsymbol{\omega} \left( \vec{\mathbf{B}}_0 \cdot \vec{\mathbf{A}} \right) \tag{31}$$



#### Figure 23

If you use an alternating current in the Helmholtz coils, then B has an alternating amplitude  $B = B_0 \cos(\omega t)$ . You can then easily map this field with the detector loop shown above. If you orient the loop so that the signal on the oscilloscope is a maximum, then you know that  $\vec{B}$  is perpendicular to the detector loop and has a magnitude given by

$$V = V_0 \sin(\omega t) = d\Phi_B/dt = d/dt \Big( NAB\cos(\omega t) \Big).$$

This is essentially the same formula we had for the rotating coil gaussmeter, Equation (27). The difference is that by "rotating the field" rather than the coil, we avoid sliding contacts, motors, electrical noise, and can make very precise measurements.

A feature of Equation (31) that we did not have when we rotated the coil is the dot product  $\vec{B}_0 \cdot \vec{A}$ . When the detector coil is aligned so that its area vector  $\vec{A}$  (which is perpendicular to the plane of the detector coil) is parallel to  $\vec{B}_{0}$ , the dot product  $\vec{B}_0 \cdot \vec{A}$  is a maximum. Thus we not only measure the magnitude of  $\vec{B}_0$ , we also get the direction by reorienting the detector coil until the V<sub>0</sub> is a maximum.

As a result, a small coil attached to an oscilloscope, which is our  $\oint \vec{E} \cdot d\vec{\ell}$  meter, can be used to accurately map the magnitude and direction of the magnetic field of the Helmholtz coils, or of any coil of wire. Unlike our earlier electric field mapping experiments, there are no mysteries or unknown constants. Faraday's law, through Equation (31), gives us a precise relation between the observed voltage and the magnetic field. The experimental setup is seen in Figure (24).

Still another way to measure magnetic fields is illustrated in Exercise 7.



#### Figure 24

Experimental setup for the magnetic field mapping experiment. A 60 cycle AC current is running through the Helmholtz coils, producing an alternating magnetic flux through the 10 turn search coil. The resulting induced voltage is seen on the oscilloscope screen.

#### Exercise 7

The point of this experiment is to determine the strength of the magnetic field produced by the small magnets that sat on the angle iron bars in the velocity detector apparatus. We placed a short piece of wood between two magnets so that there was a small gap between the ends as seen in the actual size computer scan of Figure (25). The pair of magnets were then suspended over the air track as shown in Figure (26).

On top of the air cart we mounted a single turn coil. When the air cart passes under the magnets, the single turn coil passes through the lower gap between the magnets as shown. The dimensions of the single turn coil are shown in Figure 27. We also show the dimensions and location of the lower end of one of the magnets at a time when the coil has passed part way through the gap. You can see that, at this point, all the magnetic flux across the lower gap is passing completely through the single turn coil. Figure 28 is a recording of the induced voltage in the single turn coil as the coil passes completely through the gap. The left hand blip was produced when the coil entered the gap, and the right hand blip when the coil left the gap. The air track was horizontal, so that the speed of the air cart was constant as the coil moved through the gap. Determine the strength of the magnetic field B in the gap. Show and explain your work.

Figure 25 Two C Magnets with wood spacer.





## Figure 27

Dimensions of the single turn coil. We also show the dimensions of the end of the magnets through which the coil is passing.



#### Figure 26





#### Figure 28

Voltage induced in the single turn coil.

#### **Exercise 8**

As shown in Figure (29), we started with a solenoid with 219 turns wrapped in a 1" diameter plastic tube. The coil is 45.4 cm long. The current going through the coil first goes through a .1 $\Omega$  resistor. By measuring the voltage V<sub>1</sub> across that resistor, we can determine the current through the solenoid. V<sub>1</sub> is shown as the lower curve in Figure (30).

a) Using  $V_1$  from Figure 30, calculate the magnitude B of the magnetic field in the solenoid.

We then wound 150 turns of wire around the center section of the solenoid, as indicated in Figure (29). You can see that the entire flux  $\Phi_1$  of the Magnetic field of the solenoid, goes through all the turns of the outer coil.

b) Use this fact to predict the voltage  $V_2$  across the outer coil, and then compare your prediction with the experimental  $V_2$  shown in the upper curve of Figure (30).



#### Figure 29

The inner (primary) coil 1 is 45.4 cm long, has 219 turns and is wound on a 2.54 cm (1") diameter tube. The outer (secondary) coil consists of 150 turns wound tightly around the center section of the primary coil. The current through the primary coil goes through a  $.1\Omega$  resistor, and the voltage  $V_1$  is measured across that resistor.  $V_2$  is the voltage induced in the secondary coil.



#### Figure 30

The voltage  $V_1$  across the .1 $\Omega$  resistor measures the current in the primary (219 turn) coil.  $V_2$  is the voltage induces in the secondary (outer 150 turn) coil.

# Chapter 31 Induction and Magnetic Moment

In this chapter we discuss several applications of Faraday's law and the Lorentz force law. The first is to the **inductor** which is a common electronic circuit element. We will pay particular attention to a circuit containing an inductor and a capacitor, in which an electric current oscillates back and forth between the two. Measurements of the period of the oscillation and dimensions of the circuit elements allows us to predict the speed of light without looking at light. Such a prediction leads to one of the basic questions faced by physicists around the beginning of the 20th century: who got to measure this predicted speed? The answer was provided by Einstein and his special theory of relativity.

In the second part of this chapter we will discuss the torque exerted by a magnetic field on a current loop, and introduce the concept of a **magnetic moment**. This discussion will provide some insight into how the presence of iron greatly enhances the strength of the magnetic field in an electromagnet. However the main reason for developing the concept of magnetic moment and the various magnetic moment equations is for our later discussion of the behavior of atoms and elementary particles in a magnetic field. It is useful to clearly separate the classical ideas discussed here from the quantum mechanical concepts to be developed later.

## THE INDUCTOR

In our discussion of Faraday's law and the betatron in Chapter 30, particularly in Figure (30-15), we saw that an increasing magnetic field in the core of the betatron creates a circular electric field around the core. This electric field was used to accelerate the electrons.

A more common and accessible way to produce the same circular electric field is by turning up the current in a solenoid as shown in Figure (1). As we saw in our discussion of Ampere's law in Chapter 29, a current i in a long coil of wire with n turns per unit length,



a) side view of coil and magnetic field



*b)* top view showing the electric field surrounding the increasing magnetic flux

#### Figure 1

When we turn up the current in a solenoid, we increase the magnetic field and therefore the magnetic flux up through the coil. This increasing magnetic flux is the source of the circular electric field seen in the top view. produces a nearly uniform magnetic field inside the coil whose strength is given by the formula

$$\mathbf{B} = \boldsymbol{\mu}_0 \mathbf{n} \mathbf{i} \tag{29-31}$$

and whose direction is given by the right hand rule as shown in the side view, Figure (1a).

If the coil has a cross-sectional area A, as seen in the top view Figure (1b), then the amount of magnetic flux  $\Phi_B$  "flowing" up through the coil is given by

$$\Phi_{\rm B} = BA = \mu_0 nAi \tag{1}$$

And if we are increasing the current i in the coil, then the rate of increase of this flux is (since  $\mu_0$ , n and A are constants)

$$\frac{d\Phi_{\rm B}}{dt} = \mu_0 n A \frac{di}{dt}$$
(2)

It is the changing magnetic flux that creates the circular electric field E shown in Figure (1b).



#### Figure 2 Sign conventions

We start by defining up, out of the paper, as the positive direction. Then use the right hand rule to define a positively oriented path. As a result, counter clockwise is positive, clockwise is negative. With these conventions,  $d\Phi_B/dt$  is positive for an increasing upward directed magnetic flux. In calculating the line integral  $\oint \vec{E} \cdot d\vec{\ell}$ , we go around in a positive direction, counter clockwise. Everything is positive except the – sign in Faraday's law, thus the electric field goes around in a negative direction, clockwise as shown. In Figure (2) we have shown the top view of the solenoid in Figure (1) and added in the circular electric field we would get if we had an increasing magnetic flux up through the solenoid. We have also drawn a circular path of radius r around the solenoid as shown. If we calculate the line integral  $\oint \vec{E} \cdot d\vec{\ell}$  for this closed path, we get by Faraday's law

$$\oint \vec{E} \cdot d\vec{\ell} = -\frac{d\Phi_{\rm B}}{dt}$$
(30-17)

$$E \times 2\pi r = -\mu_0 n A \frac{di}{dt}$$
(3)

where the integral  $\oint \vec{E} \cdot d\vec{\ell}$  is simply E times the circumference of the circle, and we used Equation (2) for  $d\Phi_B/dt$ .

The minus sign in Equation (3) tells us that if we use a positive path as given by the right hand rule, and we are increasing the flux up through this path, then  $\vec{E} \cdot d\vec{\ell}$  must be negative. I.e., the electric field must go clockwise, opposite to the positive path. (Do not worry too much about signs in this discussion. We will shortly find a simple, easily remembered, rule that tells us which way the electric field points.)

Equation (3) tells us that the strength E of the circular field is proportional to the rate of change of current i in the solenoid, and drops off as 1/r if we are outside the solenoid. In the following exercise, you are to show that we also have a circular field inside the solenoid, a field that decreases linearly to zero at the center.

#### Exercise 1

Use Faraday's law to calculate the electric field inside the solenoid. Note that for a circular path of radius r inside the solenoid, the flux  $\Phi_{\rm B}$  through the path is proportional to the area of the path and not the area A of the solenoid.

The calculation of the circular electric field inside and outside a solenoid, when i is changing, is a good example of the use of both Ampere's law to calculate  $\vec{B}$  and Faraday's law to calculate  $\vec{E}$ . It should be saved in your collection of good examples.

## **Direction of the Electric Field**

In Figure (2) and in the above exercise, we saw that an increasing magnetic flux in the coil created a clockwise circular electric field both inside and outside the wire as shown in Figure (3). In particular we have a circular electric field *at the wire*, and this circular electric field will act on the charges carrying the current in the wire.

To maintain our sign conventions, think of the current in the wire as being carried by the flow of positive charge. The up directed magnetic field of Figure (3) will be produced by a current flowing counterclockwise as shown (right hand rule). In order to have an increasing flux, this counterclockwise current must be increasing.

We saw that the electric field is clockwise, opposite to the direction of the current. We are turning up the current to increase the magnetic field, and the electric field is opposing the increase.

If we already have a current in a solenoid, already have an established  $\vec{B}$  field and try to decrease it, di/dt is negative for this operation, and we get an extra minus sign in Equation (3) that reverses the direction of  $\vec{E}$ . As



#### Figure 3

If the sign conventions described in Fig. 2 seemed too arbitrary, here is a physical way to determine the direction of  $\vec{E}$ . The rule is that the electric field  $\vec{E}$ opposes any change in the current i. In this case, to create an increasing upward directed magnetic flux, the current i must be flowing counter clockwise as shown, and be increasing. To oppose this increase, the electric field must be clockwise. a result we get a counterclockwise electric field that exerts a force in the direction of i. Thus when we try to decrease the current, the electric field tries to maintain it.

There is a general rule for determining the direction of the electric field. *The electric field produced by the changing magnetic flux always opposes the change*. If you have a counterclockwise current and increase it, you will get a clockwise electric field that opposes the increase. If you have a counterclockwise current and decrease it you get a counterclockwise electric field that opposes the decrease. If you have a clockwise current and try to increase it, you get a counter clockwise electric field that opposes the increase, etc. There are many possibilities, but one rule—the electric field always opposes the change.



#### Figure 4

The electric field penetrates the wire, opposing the change in the current i. The line integral  $\oint \vec{E} \cdot d\ell$  around the coil is just equal to the change in voltage  $\Delta V_1$  around each turn of the coil.

#### Induced Voltage

We have just seen that the changing magnetic flux in a solenoid creates an electric field that acts on the current in the solenoid to oppose the change in the current. From Equation (3), we see that the formula for the line integral of this electric field around one loop of the coil is given by

$$\oint \vec{E} \cdot d\vec{\ell} = -\mu_0 n A \frac{di}{dt}$$
(4)

where the path is at the wire as shown in Figure (4). The n in Equation (4), which comes from the formula for the magnetic field of a solenoid, is the number of turns per unit length in the solenoid.

In our discussion of the betatron, we saw that the circular electric field accelerated electrons as they went around the evacuated donut. Each time the electrons went around once, they gained an amount of kinetic energy which, in electron volts, was equal to  $\oint \vec{E} \cdot d\vec{\ell}$ . In our discussion of the electron gun, we saw that using a battery of voltage  $V_{acc}$  to accelerate the electrons, produced electrons whose kinetic energy, in electron volts, was equal to  $V_{acc}$ . In other words, the circular electric field can act like a battery of voltage  $V_{acc} = \oint \vec{E} \cdot d\vec{\ell}$ .

When acting on the electrons in one loop of wire, the circular electric field produces a voltage change  $\Delta V_1$  given by

$$\Delta V_1 = \oint \vec{E} \cdot d\vec{\ell} \qquad \begin{array}{c} change \ in \ electric \\ voltage \ in \ one \\ turn \ of \ the \ coil \end{array} \tag{5a}$$

If we have a coil with N turns as shown in Figure (5), then the change in voltage  $\Delta V_N$  across all N turns is N times as great, and we have

$$\Delta V_{N} = N \oint \vec{E} \cdot d\vec{\ell} \qquad \begin{array}{c} change \text{ in electric} \\ voltage \text{ in } N \\ turns \text{ of the coil} \end{array} \tag{5b}$$

Using Equation 4 for the  $\oint \vec{E} \cdot d\vec{\ell}$  for a solenoid, we see that the voltage change  $\Delta V_N$  across the entire solenoid has a magnitude

$$|\Delta \mathbf{V}_{\mathbf{N}}| = \left| \mathbf{N} \oint \vec{\mathbf{E}} \cdot d\vec{\ell} \right| = \left( \mu_0 \mathbf{N} \mathbf{n} \mathbf{A} \right) \frac{d\mathbf{i}}{d\mathbf{t}}$$
(6)

where N is the total number of turns, n = N/h is the number of turns per unit length, A is the cross-sectional area of the solenoid, and i the current through it.

To get the correct sign of  $\Delta V_N$ , to see whether we have a voltage rise or a voltage drop, we will use the rule that the circular electric field opposes any change in the current. This rule is much easier to use than trying to keep track of all the minus signs in the equations.

In summary, Equation (6) is telling us that if you try to change the amount of current flowing in a solenoid, if di/dt is not zero, then a voltage will appear across the ends of the solenoid. The voltage has a magnitude proportional to the rate di/dt that we are trying to change the current, and a direction that opposes the change. It is traditional to call this voltage  $\Delta V_N$  the *induced* voltage. One says that the changing magnetic flux in the coil *induces* a voltage. Such a coil of wire is often called an *inductor*.



#### Figure 5

Our standard coil with N turns, area A and length h. If you try to increase the current i in the coil, you get an opposing voltage.

#### Inductance

If you take a piece of insulated wire, tangle it up in any way you want, and run a current through it, you will get an induced voltage  $V_{induced}$  that is proportional to the rate of change of current di/dt, and directed in a way such that it opposes the change in the current. If we designate the proportionality constant by the letter L, then the relationship between  $V_{induced}$  and di/dt can be written

$$V_{\text{induced}} = L \frac{\text{di}}{\text{dt}} \qquad \begin{array}{c} \text{induced voltages} \\ \text{are proportional} \\ \text{to dildt} \end{array} \tag{7}$$

The constant L is called the *inductance* of the coil or tangle of wire. In the MKS system, inductance has the dimension of volt seconds/ampere, which is called a *henry*.

Comparing Equations (6) and (7), we immediately obtain the formula for the inductance of a solenoid

$$L = \mu_0 NnA = \frac{\mu_0 N^2 A}{h} \qquad inductance of \\ a \ solenoid \qquad (8)$$

where N is the number of turns in the solenoid, A the cross-sectional area and h the length. In the middle term, n = N/h is the number of turns per unit length.

#### **Example 1 The toroidal Inductor**

The simplest solenoid we can use is a toroidal one, like that shown in Figure (6), where the magnetic field is completely confined to the region inside the coil. Essentially, the toroid is an ideal solenoid (no end effects) of length  $h = 2\pi R$ . To develop an intuitive feeling for inductance and the size of a henry, let us calculate the inductance of the toroidal solenoid shown in the photograph of Figure (6b). This solenoid has 696 turns and a radius of R = 21.5 cm. Each coil has a radius of r = 1.3 cm. Thus we have

N = 696 turns  
R = 21.5 cm  
h = 
$$2\pi R$$
 =  $2\pi * .215$  = 1.35 m  
r = 1.3 cm  
A =  $\pi r^2$  =  $\pi \times .013^2$  =  $5.31 \times 10^{-4} m^2$   
 $\mu_0$  =  $1.26 \times 10^{-6}$  henry/m



**Figure 6a** A toroid is an ideal solenoid of length  $h = 2\pi R$ .

With

$$L = \frac{\mu_0 N^2 A}{h}$$

we get

$$L = \frac{1.26 \times 10^{-6} \times (696)^2 \times 5.31 \times 10^{-4}}{1.35}$$
  
= 2.40×10<sup>-4</sup> henry

We see that even a fairly big solenoid like the one shown in Figure (6) has a small inductance at least when measured in henrys. At the end of the chapter we will see that inserting an iron core into a solenoid greatly increases the inductance. Inductances as large or larger than one henry are easily obtained with iron core inductors.



#### Figure 6b

Photograph of the toroidal solenoid used in various experiments. Although the coil looks big, the inductance is only  $2.40 \times 10^{-4}$  henry. (If you put iron inside the coil, you could greatly increase the inductance, but you would not be able to calculate its value.)

## INDUCTOR AS A CIRCUIT ELEMENT

Because a changing electric current in a coil of wire produces a voltage rise, small coils are often used as circuit elements. Such a device is called an inductor, and the symbol used in circuit diagrams is a sketch of a solenoid — ())),— and usually designated by the symbol L. The voltage rise across the three circuit elements we have considered so far are

$$V_R = iR$$
 resistor (27-8)

$$V_{\rm C} = \frac{Q}{C}$$
 capacitor (27-31)

$$V_{\rm L} = L \frac{{\rm d}i}{{\rm d}t}$$
 inductor (7)

As shown in Figure (7) the direction of the rise is opposite to the current in a resistor, toward the positive charge in a capacitor, and in a direction to oppose a change in the current i in an inductor. In (c), we are showing the direction of the voltage rise for an increasing current. The voltage in the inductor is opposing an increase in the current, just as the voltage in the resistor (a) opposes the current i itself.





## The LR Circuit

We will begin our discussion of the inductor as a circuit element with the LR circuit shown in Figure (8). Although this circuit is fairly easy to analyze, it is a bit tricky to get the current i started. One way to start the current is shown in Figure (9) where we have a battery and another resistor  $R_1$  attached as shown.

When the switch of Figure (9) has been closed for a while, we have a constant current  $i_0$  that flows down out of the battery, through the resistor  $R_1$ , around up through the inductor and back to the battery. Because the current is constant,  $di_0/dt = 0$  and there is no voltage across the inductor. When we have constant "DC" currents, inductors act like short circuits. That is why the current, given the choice of going up through the inductor L or the resistor R, all goes up through L. (To say this another way, since the voltage across L is zero, the voltage  $V_R$  across R must also be zero, and the current  $i_R = V_R/R = 0$ .) Since  $R_1$  is the only thing that limits the current in Figure (9),  $i_0$  is given by

$$i_0 = \frac{V_B}{R_1} \tag{9}$$

Equation (9) tells us that we have a serious problem if we forget to include the current limiting resistor  $R_1$ .



#### Figure 8

The LR circuit. If we have a decreasing current, the voltage in the inductor opposes the decrease and creates a voltage that continues to push the current through the resistor. But, to label the voltages for Kirchoff's law, it is easier to work with positive quantities. I.e. we label the circuit as if both i and di/dt were positive. With a positive di/dt, the voltage on the inductor opposes the current, as shown.

When we open the switch of Figure (9), the battery and resistor  $R_1$  are immediately disconnected from the circuit, and we have the simple LR circuit shown in Figure (8). Everything changes instantly **except the current i in the inductor**. The inductor instantly sets up a voltage  $V_L$  to oppose any change in the current.

Figure (10) is a recording of the voltage  $V_L$  across the inductor, where the switch in Figure (9) is opened at time t = 0. Before t = 0, we have a constant current  $i_0$  and no voltage  $V_L$ . When the switch is opened the voltage jumps up to  $V_L = V_0$  and then decays exponentially just as in the RC circuit. What we want to do is apply Kirchoff's law to Figure (8) and see if we can determine the time constant for this exponential decay.



#### Figure 9

To get a current started in an LR circuit, we begin with the extra battery and resistor attached as shown. With the switch closed, in the steady state all the current  $i_0$ flows up through the inductor because it (theoretically) has no resistance. When the switch is opened, the battery is disconnected, and we are left with the RL circuit starting with an initial current  $i_0$ . If we walk around the circuit of Figure (8) in the direction of i, and add up the voltage rises we encounter, and set the sum equal to zero (Kirchoff's law), we get

$$-iR - L\frac{di}{dt} = 0$$

$$\frac{di}{dt} + \frac{R}{L}i = 0$$
(10)

Equation (10) is a simple first order differential equation for the current i. We guess from our experimental results in Figure (10) that i should be given by an exponential decay of the form

$$i = i_0 e^{-\alpha t} \tag{11}$$



#### Figure 10

Experimental recording of the voltage in an RL circuit. We see that once the switch of Fig. 9 is opened, the voltage across the inductor jumps from zero to  $V_0 = i_0 R$ . This voltage on the inductor is trying to maintain the current now that the battery is disconnected. The voltage and the current then die with an exponential decay. (For this experiment, we used the toroidal inductor of Figure 6, with  $R = 15\Omega$ ,  $R1 = 4\Omega$ , and  $V_b = 2.5$  volts.) Differentiating Equation (11) to get di/dt, we have

$$\frac{\mathrm{di}}{\mathrm{dt}} = -\alpha i_0 \mathrm{e}^{-\alpha \mathrm{t}} \tag{12}$$

and substituting (11) and (12) in Equation (10) gives

$$-\alpha i_0 e^{-\alpha t} + \frac{R}{L} i_0 e^{-\alpha t} = 0$$
(13)

In Equation (13),  $i_0$  and  $e^{-\alpha t}$  cancel and we get

$$\alpha = R/L$$

Equation (11) for i becomes

$$i = i_0 e^{-(R/L)t} = i_0 e^{-t/(L/R)}$$
  
=  $i_0 e^{-t/T}$  (14)

We see from Equation (14) that the time constant T for the decay is

$$T = \frac{L}{R} = \begin{cases} \text{time constant} \\ \text{for the decay} \\ \text{of an LR circuit} \end{cases}$$
(15)

Everything we said about exponential decays and time constants for RC circuits at the end of Chapter (27) applies to the LR circuit, except that the time constant is now L/R rather than RC.

#### Exercise 2

The LR circuit that produced the experimental results shown in Figure (10) had a resistor whose resistance R was 15 ohms. Quickly estimate the inductance L. (You should be able to make this estimate accurate to within about 10% simply by sketching a straight line on the graph of Figure 10.) Compare your result with the inductance of the toroidal solenoid discussed in Figure (6) on page 6.

## THE LC CIRCUIT

The next circuit we wish to look at is the LC circuit shown in Figure (11). All we have done is replace the resistor R in Figure (8) with a capacitor C as shown. It does not seem like much of a change, but the behavior of the circuit is very different. The exponential decays we saw in our LR and RC circuits occur because we are losing energy in the resistor R. In the LC circuit we have no resistor, no energy loss, and we will not get an exponential decay. To see what we should get, we will apply Kirchoff's law to the LC circuit and see if we can guess the solution to the resulting differential equation.

Walking clockwise around the circuit in Figure (11) and setting the sum of the voltage rises to zero, we get

$$-\frac{Q}{C} - L\frac{di}{dt} = 0$$

$$\frac{di}{dt} + \frac{Q}{LC} = 0$$
(16)

The problem we have with Equation (16) is that we have two variables, i and Q, and one equation. But we had this problem before in our analysis of the RC circuit, and solved it by noticing that the charge Q on the capacitor is related to the current i flowing into the capacitor by

$$i = \frac{dQ}{dt}$$
(17)

If we differentiate Equation (16) once with respect to time to get

$$\frac{\mathrm{d}^2 \mathrm{i}}{\mathrm{d}t^2} + \frac{1}{\mathrm{LC}}\frac{\mathrm{d}Q}{\mathrm{d}t} = 0$$



Figure 11

The LC circuit. This is the same as the LR circuit of Figure (8), except that the resistor has been replaced by a capacitor.

Finally use Equation (17) i = dQ/dt and we get the second order differential equation

$$\frac{d^2i}{dt^2} + \frac{1}{LC}i = 0$$
 (18)

The fact that we get a second order differential equation (with a second derivative of i) instead of the first order differential equations we got for LR and RC circuits, shows that we have a very different kind of problem. If we try an exponential decay in Equation (18), it will not work.

#### Exercise 3

Try the solution  $i = i_0 e^{-\alpha t}$  in Equation (18) and see what goes wrong.

We have previously seen a second order differential equation in just the form of Equation (18) in our discussion of simple harmonic motion. We expect a sinusoidal solution of the form

$$i = i_0 \sin \omega t$$
 (19)

In order to try this guess, Equation (19), we differentiate twice to get

$$\frac{di}{dt} = \omega i_0 \cos \omega t$$

$$\frac{d^2 i}{dt^2} = -\omega^2 i_0 \sin \omega t \qquad (20)$$

and substitute Equation (20) into (18) to get

$$-\omega^2 i_0 \sin \omega t + \frac{i_0}{LC} \sin \omega t = 0$$
 (21)

The quantity  $i_0 \sin \omega t$  cancels from Equation (21) and we get

$$\omega^2 = \frac{1}{LC}$$
;  $\omega = \frac{1}{\sqrt{LC}}$  (22)

We see that an oscillating current is a solution to Kirchoff's law, and that the frequency  $\omega$  of oscillation is determined by the values of L and C.

#### **Exercise 4**

In Figure (12) we have an LC circuit consisting of a toroidal coil shown in Figure (6) (on page 31-6), and the parallel plate capacitor made of two aluminum plates with small glass spacers. The voltage in Figure (12c) is oscillating at the natural frequency of the circuit.

a) What is the capacitance of the capacitor?

b) The aluminum plates have a radius of 11 cm. Assuming that we can use the parallel plate capacitor formula

$$C = \frac{\varepsilon_0 A_C}{d}$$

where  $A_C$  is the area of the plates, estimate the thickness d of the glass spacers used in this experiment.

(The measured value was 1.56 millimeters. You should get an answer closer to 1 mm. Errors could arise from fringing fields, effect of the glass, and non-uniformity of the surface of the plates.)



a) The LC circuit



b) Inductor and capacitor used in the experiment



c) Oscillating voltage at the resonant frequency.

#### Figure 12

Oscillating current in an LC circuit consisting of the toroidal inductor of Fig. 6 and a parallel plate capacitor. We will discuss shortly how we got the current oscillating and measured the voltage.

## Intuitive Picture of the LC Oscillation

The rather striking behavior of the LC circuit deserves an attempt at an intuitive explanation. The key to understanding why the current oscillates lies in understanding the behavior of the inductor. As we have mentioned, the voltage rise on an inductor is always in a direction to oppose a change in current. The closest analogy is the concept of inertia. If you have a massive object, a large force is required to accelerate it. But once you have the massive object moving, a large force is required to stop it.

An inductor effectively supplies inertia to the current flowing through it. If you have a large inductor, a lot of work is required to get the current started. But once the current is established, a lot of effort is required to stop it. In our LR circuit of Figure (10), once we got a current going through the inductor L, the current continued to flow, even though there was no battery in the circuit, because of the inertia supplied to the current by the inductor.

Let us now see why an LC circuit oscillates. One cycle of an oscillation is shown in Figure (13) where we begin in (a) with a current flowing up through the inductor and over to the capacitor. The capacitor already has some positive charge on the upper plate and the current is supplying more. The capacitor voltage  $V_C$  is opposing the flow of the current, but the inertia supplied to the current by the inductor keeps the current flowing.

In the next stage, (b), so much charge has built up in the capacitor,  $V_C$  has become so large, that the current stops flowing. Now we have a charged up capacitor which in (c) begins to discharge. The current starts to flow back down through tin inductor. The current continues to flow out of the capacitor until we reach (d) where the capacitor is finally discharged.

The important point in (d) is that, although the capacitor is empty, we still have a current and the inductor gives the current inertia. The current will continue to flow even though it is no longer being pushed by the capacitor. Now in (e), the continuing current starts to charge the capacitor up the other way. The capacitor voltage is trying to slow the current down but the inductor voltage keeps it going. Finally, in (f), enough positive charge has built up on the bottom of the capacitor to stop the flow of the current. In (g) the current reverses and the capacitor begins to discharge. The inductor supplies the inertia to keep this reversed current going until the capacitor is charged the other way in (i). But this is the same picture as (a), and the cycle begins again.

This intuitive picture allows you to make a rough estimate of how the frequency of the oscillation should depend upon the size of the inductance L and capacitance C. If the inductance L is large, the current has more inertia, it will charge up the capacitor more, and should take longer. If the capacitance C is larger, it should take longer to fill up. In other words, the period should be longer, the frequency  $\omega$  lower, if either L or C are increased. This is consistent with the result  $\omega = 1/\sqrt{LC}$  we saw in Equation (22).

Before leaving Figure (13) go back over the individual sketches and check two things. First, verify that Kirchoff's law works for each stage; i.e., that the sum of the voltage rises around the circuit is zero for each stage. Then note that whenever there is a voltage  $V_L$  on the inductor, the direction of  $V_L$  always opposes the *change* in current.



**Figure 13** The various stages in the oscillation of the electric charge in an LC circuit.

## The LC Circuit Experiment

The oscillation of the LC circuit in Figure (13) is a resonance phenomena and the frequency  $\omega_0 = 1/\sqrt{LC}$  is the resonance frequency of the circuit. If we drive the circuit, force the current to oscillate, it will do so at any frequency, but the response is biggest when we drive the circuit at the resonant frequency  $\omega_0$ .

There turns out to be a very close analogy between the LC circuit and a mass hanging on a spring as shown in Figure (14). The amplitude of the current in the circuit is analogous to the amplitude of the motion of the mass. If we oscillate the upper end of the spring at a low frequency  $\omega$  much less than the resonant frequency  $\omega_0$ , the mass just moves up and down with our hand. If  $\omega$  is much higher than  $\omega_0$ , the mass vibrates at a small amplitude and its motion is out of phase with the motion of our hand. I.e., when our hand comes down, the spring comes up, and vice versa. But when we oscillate our hand at the resonant frequency, the amplitude of vibration increases until either the mass jumps

off the spring or some form of dampening or energy loss comes into play.

It is clear from Figure (14) how to drive the motion of a mass on a spring; just oscillate our hand up and down. But how do we drive the LC circuit? It turns out that for the parallel plate capacitor and air core toroidal inductor we are using, the resonance is so delicate that if we insert something into the circuit to drive it, we kill the resonance. We need a way to drive it from the outside, and an effective way to do that is shown in Figure (15).



Run a wire from an oscillator around the coil and back to the oscillator. Do the same for the scope.





magnetic field created by the current from the oscillator

#### Figure 15

Driving the LC circuit. The turns of wire from the oscillator produce an oscillating magnetic field inside the coil. This in turn produces an electric field at the coil wires which also oscillates and drives the current in the coil. (A second wire wrapped around the coil is used to detect the voltage. The alternating magnetic field in the coil produces a voltage in the scope wire.)



#### Figure 14

To get the mass on the end of a spring oscillating at some frequency  $\omega$  you move your hand up and down at the frequency  $\omega$ . If  $\omega$  is the resonant frequency of the mass and spring system, the oscillations become quite large. In that figure we have taken a wire lead, wrapped it around the toroidal coil a couple of times, and plugged the ends into an oscillator as shown. (Some oscillators might not behave very well if you short them out this way. You may have to include a series resistor with the wire that goes to the oscillator.) When we turn on the oscillator, we get a current  $i_{osc} = i_0 \sin \omega t$  in the wire, where the frequency  $\omega$  is determined by the oscillator setting.

The important part of this setup is shown in Figure (15b) where the wire lead wraps a few times around the toroid. Since the wire lead itself forms a small coil and since it carries a current  $i_{osc}$ , it will create a magnetic field  $\vec{B}_{osc}$  as shown. Part of the field  $B_{osc}$  will lie inside the toroid and create magnetic flux  $\Phi_{osc}$  down the toroid. Since the current producing  $\vec{B}_{osc}$  is oscillating at a frequency  $\omega$ , the field and the flux will also oscillate at the same frequency. As a result we have an oscillating magnetic flux in the toroid, which by Faraday's law creates an electric field of magnitude  $\oint \vec{E} \cdot d\vec{\ell} = -d\Phi_{osc}/dt$  around the turns of the solenoid. This electric field induces a voltage in the toroid.

which drives the current in the LC circuit. We can change the driving frequency simply by adjusting the oscillator.

To detect the oscillating current in the coil, we wrap another wire around the coil, and plug that into an oscilloscope. The changing magnetic flux in the coil induces a voltage in the wire, a voltage that is detected by the scope.

In Figure (16a) we carried out the experiment shown in Figure (15), and recorded the amplitude  $V_C$  of the capacitor voltage as we changed the frequency  $\omega$  on the oscillator. We see that the amplitude is very small until we get to a narrow band of frequencies centered on  $\omega_0$ , in what is a typical resonance curve. The height of the peak at  $\omega = \omega_0$  is limited by residual resistance in the LC circuit. Theory predicts that if there were no resistance, the amplitude at  $\omega = \omega_0$  would go to infinity, but the wires in the toroid would melt first. In general, however, the less resistance in the circuit, the narrower the peak in Figure (16a), and the sharper the resonance.



Figure 16a

As we tune the oscillator frequency through the resonant frequency, the amplitude of the LC voltage goes through a peak.





**Figure 16b** *The LC apparatus.* 

## **MEASURING THE SPEED OF LIGHT**

The main reason we have focused on the LC resonance experiment shown in Figure (15) is that this apparatus can be used to measure the speed of light. We will first show how, and then discuss the philosophical implications of such a measurement.

The calculation is straightforward but a bit messy. We start with Equation (22) for the resonance frequency  $\omega_0$ 

$$\omega_0 = \frac{1}{\sqrt{LC}} \tag{22}$$

and then use Equation (8) for the inductance L of a solenoid

$$L = \mu_0 N^2 A/h \tag{8}$$

and Equation (27-32) for the capacitance of a parallel plate capacitor

$$C = \varepsilon_0 A_C / d \tag{27-32}$$

For the apparatus shown in Figure (16b), the length of the toroidal solenoid is  $h = 2\pi R_T$ , and the cross-sectional area is  $A = A_T$ , so that Equation (8) becomes

$$L_{\text{toroid}} = \mu_0 N^2 A_T / 2\pi R_T$$
(23)

For the capacitor,  $A_C$  is the area of the plates, d their separation, and we can use Equation (27-32) as it stands. If we square Equation (22) to remove the square root

$$\omega_0^2 = \frac{1}{LC}$$
(22a)

and use Equations (23) and (27-32) for L and C, we get

$$\omega_0^2 = \frac{2\pi R_T}{\mu_0 N^2 A_T} \times \frac{d}{\epsilon_0 A_C}$$
$$= \left(\frac{1}{\mu_0 \epsilon_0}\right) \times \left(\frac{2\pi R_T d}{N^2 A_T A_C}\right)$$
(24)

The important point is that the product  $\mu_0 \varepsilon_0$  appears in Equation (24), and we can solve for  $1/\mu_0 \varepsilon_0$  to get

$$\frac{1}{\mu_0 \varepsilon_0} = \frac{\omega_0^2 N^2 A_T A_C}{2\pi R_T d}$$
(25)

Finally, recall in our early discussion of magnetism, that  $\mu_0 \epsilon_0$  was related to the speed of light c by

$$c^2 = \frac{1}{\mu_0 \varepsilon_0} \tag{27-18}$$

Using Equation (25) in (27-18), and taking the square root gives

$$c = \frac{1}{\sqrt{\mu_0 \varepsilon_0}} = \omega_0 N \sqrt{\frac{A_T A_C}{2\pi R_T d}}$$
(26)

#### **Exercise 5**

Show that c in Equation (26) has the dimensions of a velocity. (Radians are really dimensionless.)

At first sight Equation (26) appears complex. But look at the quantities involved.

 $\omega_0$  = the measured resonant frequency

N = the number of turns in the solenoid

- $A_T = cross-sectional area of the toroid$
- $A_{\rm C}$  = area of capacitor plates
- $R_T$  = radius of toroid
  - d = separation of capacitor plates

Although it is a lot of stuff, everything can be counted, measured with a ruler, or in the case of  $\omega_0$ , determined from the oscilloscope trace. And the result is the speed of light c. We have determined the speed of light from a table top experiment that does not involve light.

#### **Exercise 6**

The resonant curve in Figure (16a) was measured using the apparatus shown in Figure (16b). For an inductor, we used the toroid described in Figure (6). The parallel plates have a radius of 11 cm, and a separation d =1.56mm. Use the experimental results of Figure (16a), along with the measured parameters of the toroid and parallel plates to predict the speed of light. (The result is about 20% low due to problems determining the capacitance, as we discussed in Exercise 3.) In our initial discussion of the special theory of relativity in Chapter 1, we pointed out that according to Maxwell's theory of light, the speed of light c could be predicted from a table top experiment that did not involve light. This theory, developed in 1860, predicted that light should travel at a speed  $c = 1/\sqrt{\mu_0 \epsilon_0}$ , and Maxwell knew that the product  $\mu_0 \epsilon_0$  could be determined from an experiment like the one we just described. (Different notation was used in 1860, but the ideas were the same.)

This raised the fundamental question: if you went out and actually measured the speed of a pulse of light as it passed by, would you get the predicted answer  $1/\sqrt{\mu_0\epsilon_0}$ ? If you did, that would be evidence that you were at rest. If you did not, then you could use the difference between the observed speed of the pulse and  $1/\sqrt{\mu_0\epsilon_0}$  as a measurement of your speed through space. This was the basis for the series of experiments performed by Michaelson and Morley to detect the motion of the earth. It was the basis for the rather firm conviction during the last half of the 19th century that the principle of relativity was wrong.

It was not until 1905 that Einstein resolved the problem by assuming that anyone who measured the speed of a pulse of light moving past them would get the answer  $c = 1/\sqrt{\mu_0 \epsilon_0} = 3 \times 10^8 \text{ m/s}$ , no matter how they were moving. And if everyone always got the same answer for c, then a measurement of the speed of light could not be used as a way of detecting one's own motion and violating the principle of relativity. The importance of the LC resonance experiment, of the determination of the speed of light without looking at light, is that it focuses attention on the fundamental questions that lead to Einstein's special theory of relativity.

In the next chapter we will discuss Maxwell's equations which are the grand finale of electricity theory. It was the solution of these equations that led Maxwell to his theory of light and all the interesting problems that were raised concerning the principle of relativity.

#### Exercise 7

In Figure (17a) we have an LRC series circuit driven by a sinusoidal oscillator at a frequency  $\omega$  radians/sec. The voltage V<sub>R</sub> is given by the equation

 $V_{R} = V_{R0} \cos{(\omega t)}$ 

as shown in the upper sketch of Figure (17b).

Knowing  $V_{R}$ , find the formulas and sketch the voltages for  $V_{L}$  and  $V_{C}.$  Determine the formulas for the amplitudes  $V_{L0}$  and  $V_{C0}$  in terms of  $V_{R0}$  and  $\omega$ .

Figure 17a An LRC circuit driven by a sinusoidal oscillator. The voltage  $V_R$  across the resistor is shown in Figure (17b).  $V_R$  $V_{\rm R} = V_{\rm R0} \cos(\omega t)$ Figure 17b V Knowing  $V_R$ , find the formulas and sketch the voltages for  $V_L$  and  $V_C$ .  $V_{L} = V_{L0}...$ V<sub>C</sub>

 $V_C = V_{C0}...$ 

The second half of this chapter which discusses the concept of magnetic moment, provides additional laboratory oriented applications of Faraday's law and the Lorentz force law. This topic contains essential background material for our later discussion of the behavior of atoms and elementary particles in a magnetic field, but is not required for the discussion of Maxwell's equations in the next chapter. You may wish to read through the magnetic moment discussion to get the general idea now, and worry about the details when you need them later.

## **MAGNETIC MOMENT**

We will see, using the Lorentz force law, that when a current loop (a loop of wire with a current flowing in it) is placed in a magnetic field, the field can exert a torque on the loop. This has an immediate practical application in the design of electric motors. But it also has an impact on an atomic scale. For example, iron atoms act like current loops that can be aligned by a magnetic field. This alignment itself produces a magnetic field and helps explain the magnetic properties of iron. On a still smaller scale elementary particles like the electron, proton, and neutron behave somewhat like a current loop in that a magnetic field can exert a torque on them. The phenomena related to this torque, although occurring on a subatomic scale, are surprisingly well described by the so called "classical" theory we will discuss here.

#### **Magnetic Force on a Current**

Before we consider a current loop, we will begin with a derivation of the force exerted by a straight wire carrying a current i as shown in Figure (18a). In that figure we have a positive current i flowing to the right and a uniform magnetic field  $\vec{B}$  directed down into the paper.





b) model of the neutral current as two rods of charge, the positive rod moving in the direction of the current i

#### Figure 18

Magnetic force on the moving charges when a current i is placed in a magnetic field.

In order to calculate the force exerted by  $\vec{B}$  on i, we will use our model of a current as consisting of rods of charge moving past each other as shown in Figure (18b). The rods have equal and opposite charge densities Q/l, and the positive rod is moving at a speed v to represent a positive current. The current i is the amount of charge per second carried past any crosssectional area of the wire. This is the amount of charge per meter, Q/l, times the number of meters per second, v, passing the cross-sectional area. Thus

$$i = \frac{Q}{\ell} \left( \frac{\text{coulombs}}{\text{meter}} \right) \times v \left( \frac{\text{meter}}{\text{second}} \right)$$
$$= \frac{Q}{\ell} v \left( \frac{\text{coulombs}}{\text{second}} \right)$$
(27)

In Figure (18b) we see that the downward magnetic field  $\vec{B}$  acts on the moving positive charges to produce a force  $\vec{F}_{O}$  of magnitude

$$\left| \vec{F}_{Q} \right| = Q \left| \vec{v} \times \vec{B} \right| = Q v B$$

which points toward the top of the page. The force  $\vec{f}$  on a unit length of the wire is equal to the force on one charge Q times the number of charges per unit length, which is  $1/\ell$ . Thus

force on a  
init length  
of wire
$$= f = F_Q \times \left(\frac{1}{\ell}\right) = \frac{QvB}{\ell}$$
(28)





Sideways magnetic force on a current in a magnetic field. The force per unit length  $\vec{f}$  is related to the charge per unit length  $\lambda$  by  $\vec{f} = \lambda \vec{v} \times \vec{B}$ . Since  $\lambda \vec{v}$  is the current  $\vec{i}$ , we get  $\vec{f} = \vec{i} \times \vec{B}$ .

Using Equation (27) to replace  $Qv/\ell$  by the current i in (28), and noting that  $\vec{f}$  points in the direction of  $\vec{i} \times \vec{B}$ , as shown in Figure (19), we get

 $\vec{f} = \vec{i} \times \vec{B}$  force per unit length exerted by a magnetic field  $\vec{B}$  on a current  $\vec{i}$  (29)

where  $\vec{i}$  is a vector of magnitude i pointing in the direction of the positive current.

#### Example 2

Calculate the magnetic force between two straight parallel wires separated by a distance r, carrying parallel positive currents  $i_1$  and  $i_2$  as shown in Figure (21).

#### Solution

The current  $i_1$  produces a magnetic field  $\vec{B}_1$ , which acts on  $i_2$  as shown in Figure (20) (and vice versa). Since  $\vec{B}_1$  is the field of a straight wire, it has a magnitude given by Equation (28-18) as

$$B_1 = \frac{\mu_0 i_1}{2\pi r}$$
(29-18)





b) end view showing the magnetic field of current i<sub>1</sub> exerting a magnetic force on current i<sub>2</sub>
 The force between parallel currents is attractive.

#### Figure 20

Force between two currents.

The resulting force per unit length on  $\vec{B}_2$  is

$$\vec{f} = \vec{i}_2 \times \vec{B}_1$$

which is directed in toward  $i_1$  and has a magnitude

$$\left|\vec{f}\right| = i_2 B_1 = \frac{\mu_0 i_1 i_2}{2\pi r}$$
 (30)

*Equation (30) is used in the MKS definition of the ampere and the coulomb.* In 1946 the following definition of the ampere was adopted:

The ampere is the constant current which, if maintained in two straight parallel conductors of infinite length, of negligible circular cross section, and placed 1 meter apart in a vacuum, would produce on each of these conductors a force equal to  $2 \times 10^{-7}$  newtons per meter of length.

Applying this definition to Equation (30), we set  $i_1 = i_2 = 1$  to represent one ampere currents, r = 1 to represent the one meter separation, and  $f = 2 \times 10^{-7}$  as the force per meter of length. We get

$$2 \times 10^{-7} = \frac{\mu_0}{2\pi}$$

From this we see that  $\mu_0$  is now a defined constant with the exact value

$$\mu_0 = 4\pi \times 10^{-7} \qquad by \ definition \tag{31}$$

With the above definition of the ampere, the coulomb is officially defined by the amount of charge carried by a one ampere current, per second, past a cross-sectional area of a wire .

Looking back over our derivation of the formula  $\vec{f} = \vec{i} \times \vec{B}$ , and then the above MKS definitions, we see that it is the magnetic force law  $\vec{F} = Q \vec{v} \times \vec{B}$  which now underlies the official definitions of charge and current.

## **Torque on a Current Loop**

In an easily performed experiment, we place a square loop of wire of sides ( $\ell$ ) and (w) as shown in Figure (21a), into a uniform magnetic field as shown in (21b). The loop is allowed to rotate around the axis and is now orientated at an angle  $\theta$  as seen in (21c).

If we now turn on a current i, we get an upward magnetic force proportional to  $\vec{i} \times \vec{B}$  in the section from point (1) to point (2), and a downward magnetic force proportional to  $\vec{i} \times \vec{B}$  in the section from point (3) to point (4). These two forces exert a torque about the axis of the loop, a torque that is trying to increase the angle  $\theta$ . (This torque is what turns the armature of an electric motor.)

Following our earlier right hand conventions, we will define the area  $\vec{a}$  of the loop as a vector whose magnitude is the area (a = l w) of the loop, and whose direction is given by a right hand rule for the current in the loop. Curl the fingers of your right hand in the direction of the positive current i and your thumb points in the direction of  $\vec{a}$  as shown in Figure (22).



a) a wire loop carrying a current i, free to turn on the axis





With this convention, the loop area  $\vec{A}$  points toward the upper left part of the page in Figure (21b) as shown. And we see that **the torque caused by the magnetic forces**, is trying to orient the loop so that the loop area  $\vec{a}$  is parallel to  $\vec{B}$ . This is a key result we will use often.

To calculate the magnitude of the magnetic torque, we note that the magnitude of the force on side (1)-(2) or side (3)-(4) is the force per unit length  $\vec{f} = \vec{i} \times \vec{B}$  times the length  $\ell$  of the side

$$\left|\vec{\mathbf{F}}_{1,2}\right| = \left|\vec{\mathbf{F}}_{3,4}\right| = f\ell = \mathbf{i}\mathbf{B}\ell$$

When the loop is orientated at an angle  $\theta$  as shown, then the lever arm for these forces is

lever arm 
$$=\frac{w}{2}\sin\theta$$

Since both forces are trying to turn the same way, the total torque is twice the torque produced by one force, and we have

torque = 
$$2 \times \frac{W}{2} \sin \theta \times iB$$
   
*2 times*  
*lever arm*  
*times force*

torque = 
$$iB(w\ell) \sin \theta$$
 (32)



c) magnetic force acting on tilted loop

**Figure 21** Analysis of the forces on a current loop in a magnetic field.

The final step is to convert Equation (32) into a vector equation. First recall that the vector torque  $\vec{\tau}$  is defined as

$$\vec{\tau} = \vec{r} \times \vec{F}$$

where  $\vec{r}$  and  $\vec{F}_{1,2}$  are shown in Figure (23). In the figure we see that  $\vec{r} \times \vec{F}$  and therefore  $\vec{\tau}$  points up out of the paper.

Next we note that in Equation (32),  $\theta$  is the angle between the magnetic field  $\vec{B}$  and the loop area  $\vec{a}$ . In addition, the vector cross product  $\vec{a} \times \vec{B}$  has a magnitude

$$\left|\vec{a} \times \vec{B}\right| = aB\sin\theta = (w\ell)B\sin\theta$$

and points up, in the same direction as the torque  $\vec{\tau}$ . Thus Equation (32) immediately converts to the vector equation

$$\vec{\tau} = i \vec{a} \times \vec{B}$$
(33)

where i is the current in the loop, and  $\vec{a}$  is the vector area defined by the right hand convention of Figure (22).



**Figure 22** *Right hand convention for the loop area A.* 

## **Magnetic Moment**

When you put a current loop in a magnetic field, there is no net force on the loop ( $\vec{F}_{1,2} = -\vec{F}_{3,4}$  in Figure 19b), but we do get a torque. Thus magnetic fields do not accelerate current loops, but they do turn them. In the study of the behavior of current loops, it is the torque that is important, and the torque is given by the simple formula of Equation (33).

This result can be written in an even more compact form if we define the *magnetic moment*  $\vec{\mu}$  of a current loop as the current i times the vector area  $\vec{a}$  of the loop

$$\vec{\boldsymbol{\mu}} \equiv i\vec{a} \qquad \begin{array}{c} definition \ of \\ magnetic \ moment \end{array} \tag{34}$$

With this definition, the formula for the torque on a current loop reduces to

$$\vec{\tau} = \vec{\mu} \times \vec{B}$$
(35)

Although we derived Equations (34) and (35) for a square loop, they also apply to other shapes such as round loops.



Figure 23

The torque  $\vec{\tau}_{1,2}$  exerted by the force  $\vec{F}_{1,2}$  acting on the side of the current loop. The vector  $\vec{r}$  is the lever arm of  $\vec{F}_{1,2}$  about the axis of the coil. You can see that  $\vec{r} \times \vec{F}_{1,2}$  points up out of the paper.

## **Magnetic Energy**

In Figure (24) we start with a current loop with its magnetic moment  $\mu$  aligned with the magnetic field as shown in (24a). We saw in Figure (21b) that this is the orientation towards which the magnetic force is trying to turn the loop.

If we grab the loop and rotate it around as shown in (24b) until  $\mu$  is finally orientated opposite  $\vec{B}$  as in (24c) we have to do work on the loop. We can calculate the amount of work we do rotating the loop from an angle  $\theta = 0$  to  $\theta = \pi$  using the angular analogy for the formula for work. The linear formula for work is

$$W = \int_{x_1}^{x_2} F_x dx$$
(10-19)

Replacing the linear force  $F_x$  by the angular force  $\tau$ , and the linear distances dx,  $x_1$ , and  $x_2$  by the angular distances d $\theta$ , 0 and  $\pi$ , we get

$$W = \int_{0}^{\pi} \tau d\theta$$
 (36)

If we let go of the loop, the magnetic force will try to reorient the loop back in the  $\theta = 0$  position shown in Figure (24a). We can think of the loop as falling back down to the  $\theta = 0$  position releasing all the energy we stored in it by the work we did. In the  $\theta = \pi$  position of Figure (24c) the current loop has a potential energy equal to the work we did in rotating the loop from  $\theta = 0$  to  $\theta = \pi$ .



#### Figure 24

The resting, low energy position of a current loop is with  $\bar{\mu}$  parallel to  $\bar{B}$  as shown in (a). To turn the loop the other way, we have to do work against the restoring torque  $\bar{\tau} = \bar{\mu} \times \bar{B}$  as shown in (b). The total work we do to get the loop into its high energy position (c) is  $2\mu B$ . We can think of this as magnetic potential energy that would be released if we let the loop flip back down again. We choose the zero of this potential energy half way between so that the magnetic potential energy ranges from  $+\mu B$  in (c) to  $-\mu B$  in (a). We can calculate the magnetic potential energy by evaluating the integral in Equation (36). From Equation (35) we have

$$\left| \vec{\tau} \right| = \left| \vec{\mu} \times \vec{B} \right| = \mu B \sin \theta$$

so that

$$W = \int_{0}^{\pi} \mu B \sin \theta d\theta$$
  
=  $-\mu B \cos \theta \Big|_{0}^{\pi} = 2\mu B$  (36a)

Thus the current loop in the  $\theta = \pi$  position of (24c) has an energy 2µB greater than the energy in the  $\theta = 0$ position of (24a).

It is very reasonable to define the zero of magnetic potential energy for the position  $\theta = \pi/2$ , half way between the low and high energy positions. Then the magnetic potential energy is +  $\mu$ B in the high energy position and –  $\mu$ B in the low energy position. We immediately guess that a more general formula for magnetic potential energy of the current loop is

$$\begin{array}{c} \text{magnetic potential} \\ \text{energy of a} \\ \text{current loop} \end{array} \right\} E_{\text{mag}} = -\vec{\mu} \cdot \vec{B}$$

$$(37)$$

This gives  $E_{mag} = +\mu B$  when the loop is in the high energy position with  $\vec{\mu}$  opposite  $\vec{B}$ , and  $E_{mag} = -\mu B$ in the low energy position where  $\vec{\mu}$  and  $\vec{B}$  are parallel. At an arbitrary angle  $\theta$ , Equation (37) gives  $E_{mag} = -\mu B \cos \theta$ , aresult you can obtain from equation (36a) if you integrate from  $\theta = 0$  to  $\theta = \theta$ , and adjust the zero of potential energy to be at  $\theta = \pi/2$ .

## Summary of Magnetic Moment Equations

Since we will be using the magnetic moment equations in later discussions, it will be convenient to summarize them in one place. They are a short set of surprisingly compact equations.



## Charge q in a Circular Orbit

Most applications of the concept of magnetic moment are to atoms and elementary particles. In the case of atoms, we can often picture the magnetic moment as resulting from an electron traveling in a circular orbit like that shown in Figure (25). In that figure we show a charge q traveling at a speed v in a circular orbit of radius r. Since charge is being carried around this loop, this is a current loop, where the current i is the amount of charge per second being carried past a point on the orbit. In one second the charge q goes around  $v/2\pi r$ times, therefore



#### Figure 25

A charged particle in a circular orbit acts like a current loop. Its magnetic moment turns out to be  $\mu = qvr/2$ .

Since the area of the loop is  $\pi r^2$ , we get as the formula for the magnetic moment

$$\mu = iA = q\left(\frac{v}{2\pi r}\right) * (\pi r^{2})$$

$$\mu = \frac{qvr}{2}$$

$$\max_{a \text{ charge } q \text{ traveling}} at a \text{ speed } v \text{ in } a \text{ circular orbit of radius } r$$
(38)

We can make a further refinement of Equation (38) by noting that the angular momentum J (we have already used L for inductance) of a particle of mass m traveling at a speed v in a circle of radius r has a magnitude J = mvr

More importantly,  $\vec{J}$  points perpendicular to the plane of the orbit in a right handed sense as shown in Figure (26a). This is the same direction as the magnetic moment  $\vec{\mu}$  seen in (26b), thus if we write Equation (38) in the form

$$\mu = \frac{q}{2m} (mvr) \tag{38a}$$

and use J for mvr, we can write (38a) as the vector equation

$$\vec{\mu} = \frac{q}{2m} \vec{J}$$
relation between the angular  
momentum  $\vec{J}$  and magnetic  
moment  $\vec{\mu}$  for a particle  
traveling in a circular orbit
(39)

Equation (39) is as far as we want to go in developing magnetic moment formulas using strictly classical physics. We will come back to these equations when we study the behavior of atoms in a magnetic field.



(a) angular momentum of a particle in a circular orbit



(b) magnetic moment of a charged particle in a circular orbit

#### Figure 26

Comparing the magnetic moment  $\vec{\mu}$  and angular momentum  $\vec{J}$  of a particle in a circular orbit, we see that  $\vec{\mu} = \frac{q}{\vec{J}}$ 

$$\vec{\mu} = \frac{q}{2m} \, \vec{j}$$

## IRON MAGNETS

In iron and many other elements, the atoms have a net magnetic moment due to the motion of the electrons about the nucleus. The classical picture is a small current loop consisting of a charged particle moving in a circular orbit as shown previously in Figures (25) and (26).

If a material where the atoms have a net magnetic moment is placed in an external magnetic field  $\vec{B}_{ext}$  the torque exerted by the magnetic field tends to line up the magnetic moments parallel to  $\vec{B}_{ext}$  as illustrated in Figure (27). This picture, where we show all the atomic magnetic moments aligned with  $\vec{B}_{ext}$  is an exaggeration. In most cases the thermal motion of the atoms seriously disrupts the alignment. Only at temperatures of the order of one degree above absolute zero and in external fields of the order of one tesla do we get a nearly complete alignment.

Iron and a few other elements are an exception. A small external field, of the order of 10 gauss (.001 tesla) or less, can align the magnetic moments at room temperature. This happens because neighboring atoms interact with each other to preserve the alignment in an effect called *ferromagnetism*. The theory of how this interaction takes place, and why it suddenly disappears at a certain temperature (at 1043 K for iron) has been and still is one of the challenging problems of theoretical physics. (The problem was solved by Lars Onsager for a two-dimensional array of iron atoms, but no one has yet succeeded in working out the theory for a threedimensional array.)

The behavior of iron or other ferromagnetic materials depends very much on how the substance was physically prepared, i.e., on how it was cooled from the molten mixture, what impurities are present, etc. In one extreme, it takes a fairly strong external field to align the magnetic moments, but once aligned they stay there. This preparation, called magnetically hard iron, is used for permanent magnets. In the other extreme, a small external field of a few gauss causes a major alignment which disappears when the external field is removed. This preparation called *magnetically soft* iron is used for electromagnets. Our purpose in this discussion of iron magnets is not to go over the details of how magnetic moments are aligned, what keeps them aligned or what disrupts the alignment. We will consider only the more fundamental question – what is the effect of lining up the magnetic moments in a sample of matter. What happens if we line them all up as shown in Figure (27)?

A current loop has its own magnetic field which we saw in our original discussion of magnetic field patterns and which we have reproduced here in Figure (28). This is a fairly complex field shape. (Out from the loop at distances of several loop radii, the field has the shape of what is called a "dipole" magnetic field. In certain regions earth's magnetic field has this dipole magnetic field shape.)



#### Figure 27

Ferromagnetism. When you apply an external magnetic field to a piece of "magnetically soft" iron (like a nail), the external magnetic field aligns all of the magnetic moments of the iron atoms inside the iron. The magnetic field of the current loops can be enormous compared to the external field lining them up. As a result a small external field produced say by a coil of wire, can create a strong field in the iron and we have an electromagnet. This phenomena is called ferromagnetism. When you have a large collection of aligned current loops as shown in Figure (27), the magnetic fields of each of the current loops add together to produce the magnetic field of the magnet. The magnetic field of a single current loop, shown in Figure (28), is bad enough. What kind of a mess do we get if we add up the fields of thousands, billions,  $6 \times 10^{23}$  of these current loops? The calculation seems impossibly difficult.

Ampere discovered a simple, elegant way to solve the problem. Instead of adding up the magnetic fields of each current loop, he first added up the currents using a diagram like that shown in Figure (29).

We can think of Figure (29) as the top view of the aligned current loops of Figure (27). If you look at Figure (29) for a while, you see that all the currents inside the large circle lie next to, or very close to, an equal current flowing in the opposite direction. We can say that these currents inside the big circle cancel each other. They do not carry a net charge, and therefore do not produce a net magnetic field.



**Figure 28** *Magnetic field of a current loop.* 

The cancellation is complete everywhere except at the outside surface. At this surface we essentially have a single large current loop with a current i equal to the current i in each of the little loops. It was in this way that Ampere saw that the magnetic field produced by all the small current loops packed together must be the same as the magnetic field of one big loop. What an enormous simplification!



#### Figure 29

In an iron bar magnet, the iron atoms are permanently aligned. In the cross-sectional view we are looking down on the aligned current loops of the atoms. Inside the iron, we picture the currents as cancelling, leaving a net current i (the same as the current in each loop) going around the surface of the nail. This picture of a surface current replacing the actual current loops was proposed by Ampere, and the surface current is known as an Amperian current. Now let us return to Figure (27), redrawn in (30a), where we had a collection of aligned current loops. We can think of this as a model of a magnetized iron rod where all the iron atom magnetic moments are aligned. From Figure (29) we see that one horizontal layer of these current loops can be replaced by one large loop carrying a current i that goes all the way around the iron rod. This is shown again at the top of Figure (30).

Now our rod consists of a number of layers of small loops shown in (30a). If each of these layers is replaced by a single large loop, we end up with the stack of large loops shown in (30b). But this stack of large loops is just the **same current distribution we get in a solenoid!** Thus we get the remarkable result that the vector sum of the magnetic fields of all the current loops in Figure (30a) is just the simple field of a solenoid. This is why a bar magnet and a solenoid of the same size have the same field shape, as seen in Figure (31).



a) small current loops

#### Figure 30

Ampere's picture of replacing small current loop throughout the substance by large ones on the surface.

b) large Amperian	l
currents around	
surface	

Although a bar magnet and a solenoid have the same field shape, the strength of the field in a bar magnet is usually far stronger. If the majority of the magnetic moments in an iron bar are aligned, we get a field of the order of one tesla inside the bar. To obtain comparable field strengths inside a solenoid made using copper wire, we would have to use currents so strong that the copper wire would soon heat up and perhaps melt due to electrical resistance in the copper.

## The Electromagnet

If we insert a magnetically soft iron rod into the core of a solenoid as shown in Figure (32), we have an electromagnet. It only takes a small external field to align a majority of the magnetic moments in magnetically soft iron. And when the moments are aligned, we anget fields approaching one tesla,  $10^4$  gauss, as a result. This is the principle of an electromagnet where a weak field produced by a small current in the windings produces a strong field in the iron.



#### Figure 31

Comparison of the magnetic fields of an iron magnet and a solenoid. The fields are essentially the same because the Amperian currents in a bar magnet are essentially the same as the current in the coils of a solenoid. Figure (33) is a graph showing the strength of the magnetic field inside the iron core of an electromagnet as a function of the strength of the external magnetic field produced by the windings of the solenoid. In this case a toroidal solenoid was used, the iron core is an iron ring inside the toroid, and the results in Figure (33) are for one particular sample of iron. We can get different results for different samples of iron prepared in different ways.

The vertical axis in Figure (33) shows the percentage of the maximum field  $B_{max}$  we can get in the iron.  $B_{max}$  is the "saturated" field we get when all the iron atoms magnetic moments are aligned and has a typical value of about 1 tesla. We see that a very small external field of 2 gauss brings the magnetic field up to 50% of its saturated value. Getting the other 50% is much harder. We can more or less turn on the electromagnet using a 2 gauss external field, and that not much is to be gained by using a stronger external field.



#### Figure 32

In an electromagnet, some turns of wire are wrapped around an iron bar. When a current i is turned on, the magnetic field of the turns of wire provide the external field to align the iron atoms. When the current i is shut off, and the external field disappears, the iron atoms return to a random alignment and the electromagnet shuts off. Whether the iron atoms remain aligned or not, whether we have a permanent magnet or not, depends upon the alloys (impurities) in the iron and the way the iron was cooled after casting.

#### The Iron Core Inductor

When the external field is less than 2 gauss in Figure (33), we have a more or less linear relationship shown by the dotted line between the external field and the field in the iron. In this region of the curve, for  $B_{ext} < 2$  gauss, the iron is essentially acting as a magnetic field amplifier. For this sample, a 2 gauss external field produces a 50,000 gauss magnetic field in the iron, an amplification by a factor of 25,000.

If we amplify the magnetic field in our solenoid 25,000 times, we are also amplifying the magnetic flux  $\Phi_B$  by the same factor. If we have a varying current in the solenoid, but keep Bext under 2 gauss, we will get a varying magnetic field in the iron and a varying magnetic flux  $\Phi_{\rm B}$  that is roughly proportional to the current i in the solenoid. The difference that the iron makes is that the flux  $\Phi_{\rm B}$ , and the rate of change of flux  $d\Phi_{\rm B}/dt$ will be 25,000 times larger. And so will the induced voltage in the turns of the solenoid. This means that the inductance of the solenoid is also increased by 25,000 times. If we inserted an iron ring into our air core solenoid shown in Figure (6), and the iron had the same magnetic properties as the iron sample studied in Figure (33), the inductance of our toroidal solenoid would increase 25,000 times from  $1.8 \times 10^{-4}$  henry up to about 4 henrys.



#### Figure 33

Example of a magnetization curve for magnetically soft iron. The impressive feature is that an external field of only a few gauss can produce fields in excess of xxxxx gauss inside the iron.
We can easily get large inductances from iron core inductors, but there are certain disadvantages. The curve in Figure (33) is not strictly linear, therefore the inductance has some dependence on the strength of the current in the coil. When we use an AC current in the solenoid, the iron atoms have to flip back and forth to keep their magnetic moments aligned with the AC external field. There is always some energy dissipated in the process and the iron can get hot. And if we try to go to too high a frequency, the iron atoms may not be able to flip fast enough, the magnetic field in the iron will no longer be able to follow the external field, and the amplification is lost. None of these problems is present with a air core inductor that has no iron.

## **Superconducting Magnets**

The fact that iron saturates, the fact that we can do no better than aligning all the iron atoms current loops, places a fundamental limit on the usefulness of electromagnets for producing strong magnetic fields. Instead it is necessary to return to air core solenoids or other arrangements of coils of wire, and simply use huge currents.

The problem with using copper wire for coils that produce magnetic fields stronger than 1 tesla is that such strong currents are required that even the small resistance in copper produces enough heat to melt the wire. The only solutions for copper are to use an elaborate cooling system to keep the copper from heating, or do the experiment so fast that either the copper does not have time to heat, or you do not mind if it melts. The introduction in the early 1970s of superconducting wire that could carry huge currents yet had zero electrical resistance revolutionized the design of strong field magnets. Magnets made from superconducting wire, called superconducting magnets are routinely designed to create magnetic fields of strengths up to 5 tesla. Such magnets will be used in the superconducting supercollider discussed earlier, and are now found in the magnetic resonance imaging devices in most large hospitals. The major problem with the superconducting magnets is that the superconducting wire has to be cooled by liquid helium to keep the wire in its superconducting state. And helium is a rare substance (at least on earth) that is difficult to liquefy and hard to maintain as a liquid.

In the late 1980s substances were discovered that are superconducting when immersed in liquid nitrogen, an inexpensive substance to create and maintain. So far we have not been able to make wires out of these "high temperature" superconductors that can carry the huge currents needed for big superconducting magnets. But this seems to be an engineering problem that when solved, may have a revolutionary effect not just on the design and use of superconducting magnets, but on technology in general.

## APPENDIX THE LC CIRCUIT AND FOURIER ANALYSIS

The special feature of an LC circuit, like the one shown in Figure (A1), is it's resonance at an angular frequency  $\omega_0 = 1/\sqrt{LC}$ . If you drive the circuit with an oscillator that puts out a sine wave voltage V = V<sub>0</sub>sin  $\omega$ t, the circuit will respond with a large voltage output when the driving frequency  $\omega$  equals the circuit's resonant frequency  $\omega_0$ . We saw this resonance in our discussion of the LC circuit shown in Figure (12) (p 31-11).

In Chapter 16, in our discussion of Fourier analysis, we saw that a square wave of frequency  $\omega_0$  can be constructed by adding up a series of harmonic sine waves. The first harmonic, of the form  $A_1 \sin \omega_1 t$ , has the same frequency as the square wave. For a square wave the second and all even harmonics are missing. The third harmonic is of the form  $A_3 \sin \omega_3 t$ . That is, the third harmonic's frequency is three times the frequency  $\omega_1$  of the first harmonic. The fifth harmonic is of the form  $A_5 \sin \omega_5 t$ . The amplitudes  $A_1$ ,  $A_3$ ,  $A_5$ ,... of the harmonic sine waves present in the square wave, which are shown by the vertical bars in Figure (A2), were determined by Fourier analysis. (For a square wave, there is the simple relationship  $A_3 = A_1/3$ ,  $A_5 = A_1/5$ , etc.)

The point of this lab is to demonstrate the physical reality of the harmonics in a square wave. We have seen that an LC circuit can be driven to a large amplitude resonance only when the driving frequency is equal to or close to the resonant frequency  $\omega_0 = 1/\sqrt{LC}$ . To put it another way, we can use the LC circuit to detect the presence of a sine wave of frequency  $\omega_0$  in the driving signal. If that frequency is present in the driving signal, the circuit will resonate. If it is not present, the circuit will not resonate.

Our experiment is to drive an LC circuit with a square wave, and see if the various harmonics in the square wave can each cause a resonance in the circuit. For example, if we adjust the frequency  $\omega_1$  of the square wave to equal  $\omega_0$ , then we expect the first harmonic  $A_1 \sin \omega_1 t$  to drive the circuit in resonance.



#### Figure A1

The LC circuit. For this experiment we used a fairly large commercial  $9.1 \times 10^{-3}$  henry inductor. This large inductance made the circuit more stable and less noisy than when we tried to do the experiment with the toroidal inductor of Figure (12). This inductor turned out to have an internal capacitance of  $9.9 \times 10^{-10}$ farads (990 picofarads) due to the coil windings themselves. We used this internal capacitance for the capacitor C of the circuit. (The internal capacitance was determined by measuring the resonant frequency  $\omega_0 = 1/\sqrt{LC}$  and solving for C.)

By using a large inductor L, we can attach the scope directly across the inductor, as shown, without the scope having a serious effect on the circuit. The resistance R, with  $R = 150K\Omega$ , partially isolated the LC circuit from the oscillator. This allowed the oscillator to gently drive the circuit without putting out much current and without distorting the shape of the square wave. (The oscillator could not maintain a square wave when we used the LC circuit shown in Figure (12).)

If one wants to try values of C other than the internal capacitance of the coil, one can add an external capacitor in parallel with L and C of Figure (A1). If you use an external capacitor more than about 10 times the internal capacitance, the internal capacitance of the inductor can be neglected. If we then lower the frequency  $\omega_1$  of the square wave so that  $3\omega_1 = \omega_0$ , then we expect that the third harmonic  $A_3 \sin \omega_3 t$  should drive the circuit in resonance. We should get another resonance when  $5\omega_1 = \omega_0$ , and another at  $7\omega_1 = \omega_0$ , etc. When the LC circuit is driven by a square wave, there should be a whole series of resonances, where in each case one of the harmonics has the right frequency to drive the circuit. These resonances provide direct experimental evidence that the various harmonics are physically present in the square wave, that they have energy that can drive the resonance.

In Figures A3 and A4 we look at the shape of the first few harmonics in the square wave of Figure (A2), and then watch as a square wave emerges as the harmonics are added together. (This is mostly a review of what we did back in Chapter 16.) After that, we study the resonances that occur when  $\omega_1 = \omega_0$ ,  $3\omega_1 = \omega_0$ ,  $5\omega_1 = \omega_0$ , etc. Finally we drop the square wave frequency to about  $23\omega_1 = \omega_0$ , and watch the LC circuit ring like a bell repeatedly struck by a hammer.Figure A1



#### Figure A2

Fourier analysis of a square wave. The top part of the MacScope output shows an experimental square wave. We selected one cycle of the wave, chose Fourier analysis, and see that the wave consists of a series of odd harmonics. You can see the progression of amplitudes with  $A_3 = A_1/3$ ,  $A_5 = A_1/5$ , etc. In Figure A3 we show the harmonics  $A_1 \sin \omega_1 t$ ,  $A_3 \sin \omega_3 t$ , and  $A_5 \sin \omega_5 t$ . In Figure A4 we add together the harmonics to create a square wave.



#### Figure A3

Displaying selected harmonics. Note that when you select a harmonic, you not only see the shape of the harmonic, but also see the harmonic's frequency displayed. (We highlighted this display with small rectangles.) You can see, for example, that the third harmonic is 159.5 kHz, 3 times the fundamental frequency of 53.19 kHz.



#### Figure A4

Adding up harmonics to create a square wave. The more harmonics we add, the closer we get.



Figure A5 – Resonance at  $\omega_1 = \omega_0$ 

We get the biggest resonance when the frequency  $\omega_1$  of the square wave is equal to the resonant frequency  $\omega_0$  of the circuit. We displayed the first harmonic by clicking on the bar over the 1 in the Fourier analysis window, and see that the frequency of the first harmonic is 52.91 kilohertz.









Lowering the frequency of the square wave to 10.25 kilohertz, we get another resonance. In the Fourier analysis window, we selected the fifth harmonic, and see that the frequency of this harmonic is 52.63 kilohertz. To within experimental accuracy, this is again equal to the LC circuit's resonant frequency of 52.91 kilohertz.







#### Figure A9 – Ringing like a bell

Dropping the square wave frequency even further, we see that every time the voltage of the square wave changes, the circuit responds like a bell struck by a hammer. This setup can be used as the starting point for the study of damped resonant (LRC) circuits.

# Chapter 32 Maxwell's Equations

In 1860 James Clerk Maxwell summarized the entire content of the theory of electricity and magnetism in a few short equations. In this chapter we will review these equations and investigate some of the predictions one can make when the entire theory is available.

What does a complete theory of electricity and magnetism involve? We have to fully specify the electric field  $\vec{E}$ , the magnetic field  $\vec{B}$ , and describe what effect the fields have when they interact with matter.

The interaction is described by the Lorentz force law

$$\vec{F} = q\vec{E} + q \vec{v} \times \vec{B} \tag{28-18}$$

which tells us the force exerted on a charge q by the  $\vec{E}$ and  $\vec{B}$  fields. As long as we stay away from the atomic world where quantum mechanics dominates, then the Lorentz force law combined with Newton's second law fully explains the behavior of charges in the presence of electric and magnetic fields, whatever the origin of the fields may be.

To handle the electric and magnetic fields, recall our discussion in Chapter 30 (on two kinds of fields) where we saw that any vector field can be separated into two parts; a divergent part like the electric field of static charges, and a solenoidal part like the electric field in a betatron or inductor. To completely specify a vector field, we need two equations – one involving a surface integral or its equivalent to define the divergent part of the field, and another involving a line integral or its equivalent defining the solenoidal part.

In electricity theory we have two vector fields  $\vec{E}$  and  $\vec{B}$ , and two equations are needed to define each field. Therefore the total number of equations required must be four.

How many of the required equations have we discussed so far? We have Gauss' law for the divergent part of  $\vec{E}$ , and Faraday's law for the solenoidal part. It appears that we already have a complete theory of the electric field, and we do. Gauss' law and Faraday's law are two of the four equations needed.

For magnetism, we have Ampere's law that defines the solenoidal part of  $\vec{B}$ . But we have not written an equation involving the surface integral of  $\vec{B}$ . We are missing a Gauss' law type equation for the magnetic field. It would appear that the missing Gauss' law for  $\vec{B}$ , plus Ampere's law make up the remaining two equations.

This is not quite correct. The missing Gauss' law is one of the needed equations for  $\vec{B}$ , and it is easily written down because there are no known sources for a divergent  $\vec{B}$  field. But Ampere's law, in the form we have been using

$$\oint \vec{B} \cdot \vec{dl} = \mu_0 i \tag{29-18}$$

has a logical flaw that was discovered by Maxwell. When Maxwell corrected this flaw by adding another source term to the right side of Equation (29-18), he then had the complete, correct set of four equations for  $\vec{E}$  and  $\vec{B}$ . All Maxwell did was to add one term to the four equations for  $\vec{E}$  and  $\vec{B}$ , and yet the entire set of equations are named after him. The reason for this is that with the correct set of equations, Maxwell was able to obtain solutions of the four equations, predictions of these equations that could not be obtained until Ampere's law had been corrected. The most famous of these predictions was that a certain structure of electric and magnetic fields could travel through empty space at a speed  $v = 1/\sqrt{\mu_0 \varepsilon_0}$ . Since Maxwell knew that  $1/\sqrt{\mu_0 \varepsilon_0}$  was close to the observed speed  $3 \times 10^8$  m/s for light, he proposed that this structure of electric and magnetic fields was light itself.

In this chapter, we will first describe the missing Gauss' law for magnetic fields, then correct Ampere's law to get the complete set of Maxwell's four equations. We will then solve these equations for a structure of electric and magnetic fields that moves through empty space at a speed  $v = 1/\sqrt{\mu_0 \varepsilon_0}$ . We will see that this structure explains various properties of light waves, radio waves, and other components of the electromagnetic spectrum. We will find, for example, that we can detect radio waves by using the same equipment and procedures we have used in earlier chapters to detect and map electric and magnetic fields.

## GAUSS' LAW FOR MAGNETIC FIELDS

Let us review a calculation we have done several times now—the use of Gauss' law to calculate the electric field of a point particle. Our latest form of the law is

$$\int_{\text{closed}} \vec{E} \cdot d\vec{A} = \frac{Q_{\text{in}}}{\varepsilon_0}$$
(29-5)

where  $Q_{in}$  is the total amount of electric charge inside the surface.

In Figure (1), we have a point charge Q and have constructed a closed spherical surface of radius r centered on the charge. For this surface,  $\vec{E}$  is everywhere perpendicular to the surface or parallel to every surface element  $d\vec{A}$ , thus  $\vec{E} \cdot d\vec{A} = EdA$ . Since E is of constant magnitude, we get

$$\int_{closed} \vec{E} \cdot d\vec{A} = E \int_{closed} dA$$

$$= E 4 \pi r^{2}$$

$$= \frac{Q_{in}}{\varepsilon_{0}}$$
where  $Q_{in} = Q$ .



The solution of Equation (1) gives  $E = Q/4\pi\epsilon_0 r^2$  as the strength of the electric field of a point charge. A similar calculation using a cylindrical surface gave us the electric field of a charged rod. By being clever, or working very hard, one can use Gauss' law in the form of Equation (29-5) to solve for the electric field of any static distribution of electric charge.

But the simple example of the field of a point charge illustrates the point we wish to make. Gauss' law determines the diverging kind of field we get from a point source. Electric fields have point sources, namely electric charges, and it is these sources in the form of Q<sub>in</sub> that appear on the right hand side of Equation (29-5).

Figure (2) shows a magnetic field emerging from a point source of magnetism. Such a point source of magnetism is given the name *magnetic monopole* and magnetic monopoles are predicted to exist by various recent theories of elementary particles. These theories are designed to unify three of the four basic interactions – the electrical, the weak, and the nuclear interactions. (They are called "Grand Unified Theories" or "GUT" theories. Gravity raises problems that are not handled by GUT theories.) These theories also predict that the proton should decay with a half life of  $10^{32}$  years.



**Figure 2** *Magnetic field produced by a point source.* 

In the last 20 years there has been an extensive search for evidence for the decay of protons or the existence of magnetic monopoles. So far we have found no evidence for either. (You do not have to wait  $10^{32}$  years to see if protons decay; instead you can see if one out of  $10^{32}$  protons decays in one year.)

The failure to find the magnetic monopole, the fact that no one has yet seen a magnetic field with the shape shown in Figure (2), can be stated mathematically by writing a form of Gauss' law for magnetic fields with the magnetic charge  $Q_{in}$  set to zero

$$\int_{\text{closed}} \vec{B} \cdot d\vec{A} = 0$$
 (2)

When reading Equation (2), interpret the zero on the right side of Equation (2) as a statement that the divergent part of the magnetic field has no *source term*. This is in contrast to Gauss' law for electric fields, where  $Q_{in}/\epsilon_0$  is the source term.

## MAXWELL'S CORRECTION TO AMPERE'S LAW

As we mentioned in the introduction, Maxwell detected a logical flaw in Ampere's law which, when corrected, gave him the complete set of equations for the electric and magnetic fields. With the complete set of equations, Maxwell was able to obtain a theory of light. No theory of light could be obtained without the correction.

Ampere's law, Equation (29-18), uses the line integral to detect the solenoidal component of the magnetic field. We had

$$\oint \vec{B} \cdot d\vec{\ell} = \mu_0 i_{\text{enclosed}} \qquad \begin{array}{l} Ampere's \\ Law \end{array}$$
(29-18)

where  $i_{enclosed}$  is the total current encircled by the closed path used to evaluate  $\vec{B} \cdot d\vec{\ell}$ . We can say that  $\mu_0 i_{enclosed}$  is the source term for this equation, in analogy to  $Q_{in}/\epsilon_0$  being the source term for Gauss' law.

Before we discuss Maxwell's correction, let us review the use of Equation (29-18) to calculate the magnetic field of a straight current i as shown in Figure (3). In (3a) we see the wire carrying the current, and in (3b) we show the circular magnetic field produced by the current. To apply Equation (29-18) we draw a closed circular path of radius r around the wire, centering the



Using Ampere's law.

path on the wire so that  $\vec{B}$  and sections  $d\vec{l}$  of the path are everywhere parallel. Thus  $\vec{B} \cdot d\vec{l} = B dl$ , and since B is constant along the path, we have

$$\oint \vec{B} \cdot d\vec{\ell} = B \oint d\ell = B \times 2\pi r = \mu_0 i$$

which gives our old formula  $B = \mu_0 i/2\pi r$  for the magnetic field of a wire.

To see the flaw with Ampere's law, consider a circuit where a capacitor is being charged up by a current i as shown in Figure (4). When a capacitor becomes charged, one plate becomes positively charged and the other negatively charged as shown. We can think of the capacitor being charged because a positive current is flowing into the left plate, making that plate positive, and a positive current is flowing out of the right plate, making that plate negative.

Figure (4) looks somewhat peculiar in that the current i almost appears to be flowing through the capacitor. We have a current i on the left, which continues on the right, with a break between the capacitor plates. To emphasize the peculiar nature of this discontinuity in the current, imagine that the wires leading to the capacitor are huge wires, and that the capacitor plates are just the ends of the wires as shown in Figure (5).

Now let us apply Ampere's law to the situation shown in Figure (5). We have drawn three paths, Path (1) around the wire leading into the positive plate of the capacitor, Path (2) around the wire leading out of the negative plate, and Path (3) around the gap between the plates. Applying Ampere's law we have





**Figure 4** *Charging up a capacitor.* 

$$\oint_{\text{path 1}} \vec{B} \cdot d\vec{\ell} = \mu_0 i \qquad \begin{array}{c} path \ 1 \ goes \\ around \ a \ current \ i \end{array} (3)$$

$$\oint_{\text{path 2}} \vec{B} \cdot d\vec{\ell} = \mu_0 i \qquad \begin{array}{c} path \, 2 \, goes \\ around \, a \, current \, i \end{array} \tag{4}$$

$$\oint_{\text{path 3}} \vec{B} \cdot d\vec{\ell} = 0 \qquad \begin{array}{c} path \ 3 \ does \ not \ go}{around \ any \ current} \quad (5)$$

When we write out Ampere's law this way, the discontinuity in the current at the capacitor plates looks a bit more disturbing.

For greater emphasis of the problem, imagine that the gap in Figure (5) is very narrow, like Figure (5a) only worse. Assume we have a 1 mm diameter wire and the gap is only 10 atomic diameters. Then according to Ampere's law,  $\oint \vec{B} \cdot d\vec{\ell}$  should still be zero if it is correctly centered on the gap. But can we possibly center a path on a gap that is only 10 atomic diameters wide? And even if we could, would  $\oint \vec{B} \cdot d\vec{\ell}$  be zero for this path, and have the full value  $\mu_0$  i for the path 10 atomic diameters away? No, we simply cannot have such a discontinuity in the magnetic field and there must be something wrong with Ampere's law. This was the problem recognized by Maxwell.



Current flows through Paths (1) & (2), but not through Path (3).



Very narrow gap

Maxwell's solution was that even inside the gap at the capacitor there was a source for  $\oint \vec{B} \cdot d\vec{\ell}$ , and that the strength of the source was still  $\mu_0 i$ . What actually exists inside the gap is the electric field  $\vec{E}$  due to the + and – charge accumulating on the capacitor plates as shown in Figure (6). Perhaps this electric field can somehow replace the missing current in the gap.

The capacitor plates or rod ends in Figure (6) have a charge density  $\sigma = Q/A$  where Q is the present charge on the capacitor and A is the area of the plates. In one of our early Gauss' law calculations we saw that a charge density on a conducting surface produces an electric field of strength  $E = \sigma/\epsilon_0$ , thus E between the plates is related to the charge Q on them by

$$E = \frac{\sigma}{\varepsilon_0} = \frac{Q}{\varepsilon_0 A}$$
;  $Q = \varepsilon_0 E A$  (6)

The current flowing into the capacitor plates is related to the charge Q that has accumulated by

$$i = \frac{dQ}{dt}$$
(7)

Using Equation (6) in Equation (7), we get

$$i = \varepsilon_0 \frac{d}{dt} (EA)$$
(8)



**Figure 6** An electric field  $\vec{E}$  exists between the plates.

Noting that the flux  $\Phi_E$  of electric field between the plates is

$$\Phi_{\rm E} = \vec{\rm E} \cdot d\vec{\rm A} = {\rm E} {\rm A}_{\perp} = {\rm E} {\rm A}$$
(9)

and multiplying through by  $\mu_0$ , we can write Equation (8) in the form

$$\mu_0 i = \mu_0 \varepsilon_0 \frac{d\Phi_E}{dt}$$
(10)

We get the somewhat surprising result that  $\mu_0 \epsilon_0$  times the rate of change of electric flux inside the capacitor has the same magnitude as  $\mu_0 i$ , where i is the current in the wire leading to the capacitor. Maxwell proposed that  $\mu_0 \epsilon_0 d\Phi_E/dt$  played the same role, inside the capacitor, as a source term for  $\oint \vec{B} \cdot \vec{dl}$ , that  $\mu_0 i$  did outside in the wire.

As a result, Maxwell proposed that Ampere's law be corrected to read

$$\oint \vec{B} \cdot d\vec{\ell} = \mu_0 i + \mu_0 \varepsilon_0 \frac{d\Phi_E}{dt} \begin{vmatrix} corrected \\ Ampere's \\ law \end{vmatrix}$$
(11)

Applying Equation (11) to the three paths shown in Figure (7), we have

$$\oint_{\text{paths } 1\&2} \vec{B} \cdot d\vec{\ell} = \mu_0 i \qquad (\Phi_E = 0) \qquad (12)$$

$$\oint_{\text{paths 3}} \vec{B} \cdot d\vec{\ell} = \mu_0 \varepsilon_0 \frac{d\Phi_E}{dt} \quad (i=0)$$
(13)



**Figure 7** Path (2) surrounds a changing electric flux. Inside the gap,  $\mu_0 \varepsilon_0 (d\Phi_F/dt)$  replaces  $\mu_0 i$  as the source of  $\vec{B}$ . I.e., for Paths (1) and (2), there is no electric flux through the path and the source of  $\oint \vec{B} \cdot d\vec{l}$  is the current. For Path (3), no current flows through the path and the source of  $\oint \vec{B} \cdot d\vec{l}$  is the changing electric flux. But, because  $\mu_0 i$  in Equation (12) has the same magnitude  $\mu_0 \epsilon_0 d\Phi_E/dt$  in Equation (13), the term  $\oint \vec{B} \cdot d\vec{l}$  has the same value for Path (3) as (1) and (2), and there is no discontinuity in the magnetic field.

## Example: Magnetic Field between the Capacitor Plates

As an example of the use of the new term in the corrected Ampere's law, let us calculate the magnetic field in the region between the capacitor plates. To do this we draw a centered circular path of radius r smaller than the capacitor radius R as shown in Figure (8). There is no current through this path, but there is an electric flux  $\Phi_{\rm E}(r) = {\rm E}A(r) = {\rm E}\pi r^2$  through the path. Thus we set i = 0 in Ampere's corrected law, and replace  $\Phi_{\rm E}$  by the flux  $\Phi_{\rm E}(r)$  through our path to get

$$\oint \vec{B} \cdot d\vec{\ell} = \mu_0 \varepsilon_0 \frac{d\Phi_E(r)}{dt}$$
(14)

Equation (14) tells us that because we have an increasing electric field between the plates, and thus an increasing electric flux through our path, there must be a magnetic field around the path.



**Figure 8** Calculating  $\vec{B}$  in the region between the plates.

Due to the cylindrical symmetry of the problem, the only possible shape for the magnetic field inside the capacitor is circular, just like the field outside. This circular field and our path are shown in the end view, Figure (9). Since B and  $d\vec{\ell}$  are parallel for all the steps around the circular path, we have  $\vec{B} \cdot d\vec{\ell} = B d\ell$ . And since B is constant in magnitude along the path, we get

$$\oint \vec{B} \cdot d\vec{\ell} = B \oint d\ell = B \times 2\pi r$$
 (15)

To evaluate the right hand side of Equation (14), note that the flux through our path  $\Phi_E(r)$  is equal to the total flux  $\Phi_E(\text{total})$  times the ratio of the area  $\pi r^2$  of our path to the total area  $\pi R^2$  of the capacitor plates

$$\Phi_{\rm E}(\mathbf{r}) = \Phi_{\rm E}(\text{total}) \frac{\pi r^2}{\pi R^2}$$
(16)

so that the right hand side becomes

$$\mu_0 \varepsilon_0 \frac{d\Phi_E(\mathbf{r})}{dt} = \mu_0 \varepsilon_0 \frac{d\Phi_E(\text{total})}{dt} \frac{r^2}{R^2}$$
$$= \mu_0 i \frac{r^2}{R^2}$$
(17)

where in the last step we used Equation (10) to replace  $\mu_0 \epsilon_0 d\Phi_E$ (total)/dt by a term of the same magnitude, namely  $\mu_0 i$ .



End view of capacitor plate.

Finally using Equation (15) and (17) in (14) we get

$$B \times 2\pi r = \mu_0 i \frac{r^2}{R^2}$$
$$B = \frac{\mu_0 i}{2\pi} \left(\frac{r}{R^2}\right) \qquad \begin{array}{l} magnetic \\ field \ between \\ capacitor \ plates \end{array} \tag{18}$$

Figure (10) is a graph of the magnitude of B both inside and outside the plates. They match up at r = R, and the field strength decreases linearly to zero inside the plates.

#### Exercise 1

Calculate the magnetic field inside the copper wires that lead to the capacitor plates of Figure (5). Use Ampere's law and a circular path of radius r inside the copper as shown in Figure (11). Assuming that there is a uniform current density in the wire, you should get Equation (18) as an answer. Thus the magnetic field is continuous as we go out from the copper to between the capacitor plates.



Figure 11



**Figure 10** *Magnetic fields inside and outside the gap.* 

## **MAXWELL'S EQUATIONS**

Now that we have corrected Ampere's law, we are ready to write the four equations that completely govern the behavior of classical electric and magnetic fields. They are

(a) $\int_{\text{closed}} \vec{E} \cdot d\vec{A} = \frac{Q_{\text{in}}}{\varepsilon_0}$	Gauss' Law
(b) $\int_{\text{closed}} \vec{B} \cdot d\vec{A} = 0$	No Monopole
(c) $\oint \vec{B} \cdot d\vec{\ell} = \mu_0 i + \mu_0 \epsilon_0 \frac{d\Phi_E}{dt}$	Ampere's Law
(d) $\oint \vec{E} \cdot d\vec{\ell} = -\frac{d\Phi_{B}}{dt}$	Faraday's Law (19)

The only other thing you need for the classical theory of electromagnetism is the Lorentz force law and Newton's second law to calculate the effect of electric and magnetic fields on charged particles.

$$\vec{F} = q\vec{E} + q\vec{v}\times\vec{B}$$
*Lorentz Force Law*
(20)

This is a complete formal summary of everything we have learned in the past ten chapters.

## Exercise 2

This is one of the most important exercises in the text. The four Maxwell's equations and the Lorentz force law represent an elegant summary of many ideas. But these equations are nothing but hen scratchings on a piece of paper if you do not have a clear idea of how each term is used. The best way to give these equations meaning is to know inside out at least one specific example that illustrates the use of each term in the equations. For Gauss' law, we have emphasized the calculation of the electric field of a point and a line charge. We have the nonexistence of the divergent magnetic field in Figure (2) to illustrate Gauss' law for magnetic fields. We have used Ampere's law to calculate the magnetic field of a wire and a solenoid. The new term in Ampere's law was used to calculate the magnetic field inside a parallel plate capacitor that is being charged up.

Faraday's law has numerous applications including the air cart speed meter, the betatron, the AC voltage generator, and the inductance of a solenoid. Perhaps the most important concept with Faraday's law is that  $\oint \vec{E} \cdot \vec{d} \vec{l}$  is the voltage rise created by solenoidal electric fields, which for circuits can be read directly by a voltmeter. This lead to the interpretation of a loop of wire with a voltmeter attached as an  $\oint \vec{E} \cdot \vec{d} \vec{l}$  meter in the design of the air cart speed detector and experiment where we mapped the magnetic field of a Helmholtz coil.

Then there is the Lorentz force law with the formulas for the electric and magnetic force on a charged particle. As an example of an electric force we calculated the trajectory of an electron beam between charged plates, and for a magnetic force we studied the circular motion of electrons in a uniform magnetic field.

The assignment of this exercise is to write out Maxwell's equations one by one, and with each equation write down a fully worked out example of the use of each term. Do this neatly, and save it for later reference. This is what turns the hen scratchings shown on the previous page into a meaningful theory. When you buy a T-shirt with Maxwell's equations on it, you will be able to wear it with confidence.

We have just crossed what you might call a continental divide in our study of the theory of electricity and magnetism. We spent the last ten chapters building up to Maxwell's equations. Now we descend into applications of the theory. We will focus on applications and discussions that would not have made sense until we had the complete set of equations—discussions on the symmetry of the equations and applications like Maxwell's theory of light.

## SYMMETRY OF MAXWELL'S EQUATIONS

Maxwell's Equations (19 a, b, c, and d), display considerable symmetry, and a special lack of symmetry. But the symmetry or lack of it is clouded by our choice of the MKS units with its historical constants  $\mu_0$  and  $\epsilon_0$  that appear, somewhat randomly, either in the numerator or denominator at various places.

For this section, let us use a special set of units where the constants  $\mu_0$  and  $\epsilon_0$  have the value 1

$$\mu_0 = 1; \quad \varepsilon_0 = 1 \quad \left( \begin{array}{c} \text{in a special} \\ \text{set of units} \end{array} \right)$$
(21)

Because the speed of light c is related to  $\mu_0$  and  $\epsilon_0$  by  $c = 1/\sqrt{\mu_0\epsilon_0}$ , we are now using a set of units where the speed of light is 1. If we set  $\mu_0 = \epsilon_0 = 1$  in Equations (19) we get

$$\int_{\text{closed}} \vec{E} \cdot d\vec{A} = Q_{\text{in}}$$
(22a)

$$\int_{\text{closed}} \vec{B} \cdot d\vec{A} = 0$$
(22b)

$$\oint \vec{B} \cdot d\vec{\ell} = i + \frac{d\Phi_E}{dt}$$
(22c)

$$\oint \vec{E} \cdot d\vec{\ell} = -\frac{d\Phi_{\rm B}}{dt}$$
(22d)

Stripping out  $\mu_0$  and  $\varepsilon_0$  gives a clearer picture of what Maxwell's equations are trying to say. Equation (22a) tells us that electric charge is the source of divergent electric fields. Equation (22b) says that we haven't found any source for divergent magnetic fields. Equation (22c) tells us that an electric current or a changing electric flux is a source for solenoidal magnetic fields, and (22d) tells us that a changing magnetic flux creates a solenoidal electric field. Equations (22) immediately demonstrate the lack of symmetry caused by the absence of magnetic monopoles, and so does the Lorentz force law of Equation (20). If the magnetic monopole is discovered, and we assign to it the "magnetic charge"  $Q_B$ , then for example Equation (22b) would become

$$\int_{\text{closed}} \vec{B} \cdot d\vec{A} = Q_B \qquad (22b')$$

If we have magnetic monopoles, a magnetic field should exert a force  $\vec{F}_B = Q_B \vec{B}$  and perhaps an electric field should exert a force something like  $\vec{F}_E = Q_B \vec{v} \times \vec{E}$ .

Aside from Equation (22b), the other glaring asymmetry is the presence of an electric current i in Ampere's law (22c) but no current term in Faraday's law (22d). If, however, we have magnetic monopoles we can also have a current  $i_B$  of magnetic monopoles, and this asymmetry can be removed.

#### Exercise 3

Assume that the magnetic monopole has been discovered, and that we now have magnetic charge  $Q_B$  and a current  $i_B$  of magnetic charge. Correct Maxwell's Equations (22) and the Lorentz force law (20) to include the magnetic monopole. For each new term you add to these equations, provide a worked-out example of its use.

In this exercise, use symmetry to guess what terms should be added. If you want to go beyond what we are asking for in this exercise, you can start with the formula  $\vec{F}_B = Q_B \vec{B}$  for the magnetic force on a magnetic charge, and with the kind of thought experiments we used in the chapter on magnetism, derive the formula for the electric force on a magnetic charge  $Q_B$ . You will also end up with a derivation of the correction to Faraday's law caused by a current of magnetic charge. (This is more of a project than an exercise.)

## MAXWELL'S EQUATIONS IN EMPTY SPACE

In the remainder of this chapter we will discuss the behavior of electric and magnetic fields in empty space where there are no charges or currents. A few chapters ago, there would not have been much point in such a discussion, for electric fields were produced by charges, magnetic fields by currents, and without charges and currents, we had no fields. But with Faraday's law, we see that a changing magnetic field. And with the correction to Ampere's law, we see that a changing electric field. And with the correction to Ampere's law, we see that a changing electric field. Even without charges and currents we have sources for both electric and magnetic fields.

First note that if we have no electric charge (or magnetic monopoles), then we have no sources for either a divergent electric or divergent magnetic field. In empty space diverging fields do not play an important role and we can focus our attention on the equations for the solenoidal magnetic and solenoidal electric field, namely Ampere's and Faraday's laws.

Setting i = 0 in Equation (19c), the Equations (19c) and (19d) for the solenoidal fields in empty space become

$$\oint \vec{B} \cdot d\vec{\ell} = \mu_0 \varepsilon_0 \frac{d\Phi_E}{dt}$$
(23a)

$$\oint \vec{E} \cdot d\vec{\ell} = -\frac{d\Phi_{\rm B}}{dt}$$
(23b)

We can make these equations look better if we write  $\mu_0 \epsilon_0$  as  $1/c^2$ , where  $c = 3 \times 10^8$  m/s as determined in our LC circuit experiment. Then Equations (23) become

$$\oint \vec{B} \cdot d\vec{\ell} = \frac{1}{c^2} \frac{d\Phi_E}{dt} \qquad Maxwells \qquad (24a)$$

$$\oint \vec{E} \cdot d\vec{\ell} = -\frac{d\Phi_B}{dt} \qquad space \qquad (24b)$$

Equations (24a, b) suggest a coupling between electric and magnetic fields. Let us first discuss this coupling in a qualitative, somewhat sloppy way, and then work out explicit examples to see precisely what is happening.

Roughly speaking, Equation (24a) tells us that a changing electric flux or field creates a magnetic field, and (24b) tells us that a changing magnetic field creates an electric field. These fields interact, and in some sense support each other.

If we were experts in integral and differential equations, we would look at Equations (24) and say, "Oh, yes, this is just one form of the standard wave equation. The solution is a wave of electric and magnetic fields traveling through space." Maxwell was able to do this, and solve Equations (24) for both the structure and the speed of the wave. The speed turns out to be c, and he guessed that the wave was light.

Because the reader is not expected to be an expert in integral and differential equations, we will go slower, working out specific examples to see what kind of structures and behavior we do get from Equations (24). We are just beginning to touch upon the enormous subject of electromagnetic radiation.

## **A Radiated Electromagnetic Pulse**

We will solve Equations (24) the same way we have been solving all equations involving derivatives or integrals—by guessing and checking. The rules of the game are as follows. Guess a solution, then apply Equations (24) to your guess in every possible way you can think of. If you cannot find an inconsistency, your guess may be correct.

In order to guess a solution, we want to pick an example that we know as much as possible about and use every insight we can to improve our chances of getting the right answer. Since we are already familiar with the fields associated with a current in a wire, we will focus on that situation. Explicitly, we will consider what happens, what kind of fields we get, when we first turn on a current in a wire. We will see that a structure of magnetic and electric fields travels out from the wire, in what will be an example of a radiated electromagnetic pulse or wave.

#### A Thought Experiment

Let us picture a very long, straight, copper wire with no current in it. At time t = 0 we start an upward directed current i everywhere in the wire as shown in Figure (12). This is the tricky part of the experiment, having the current i start everywhere at the same time. If we closed a switch, the motion of charge would begin at the switch and advance down the wire. To avoid this, imagine that we have many observers with synchronized watches, and they all reach into the wire and start the positive charge moving at t=0. However you want to picture it, just make sure that there is no current in the wire before t = 0, and that we have a uniform current i afterward.

In our previous discussions, we saw that a current i in a straight wire produced a circular magnetic field of magnitude  $B = \mu_0 i/2\pi r$  everywhere outside the wire. This cannot be the solution we need because it implies that as soon as the current is turned on, we have a magnetic field throughout all of space. The existence of the magnetic field carries the information that we have turned on the current. Thus the instantaneous spread of the field throughout space carries this information faster than the speed of light and violates the principle of causality. As we saw in Chapter 1, we could get answers to questions that have not yet been asked.

Using our knowledge of special relativity as a guide, we suspect that the solution  $B = \mu_0 i/2\pi r$  everywhere in space, instantaneously, is not a good guess. A more reasonable guess is that the magnetic field grows at some speed v out from the wire. Inside the growing front, the field may be somewhat like its final form  $B = \mu_0 i/2\pi r$ , but outside we will assume B = 0.

Figure 12 A current i is started all along the wire at time t = 0. The pure, expanding magnetic field shown in Figure (13) seems like a good guess. But it is wrong, as we can see if we apply Ampere's law to Path (a) which has not yet been reached by the growing magnetic field. For this path that lies outside the magnetic field,  $\oint \vec{B} \cdot d\vec{l} = 0$ , and the corrected Ampere's law, Equation (19c), gives

$$\oint \vec{B} \cdot d\vec{\ell} = \mu_0 i + \mu_0 \varepsilon_0 \frac{d\Phi_E}{dt} = 0$$
 (25)

In our picture of Figure (13) we have no electric field, therefore  $\Phi_E = 0$  and Equation (25) implies that  $\mu_0 i$  is zero, or the current i through Path (a) is zero. But **the current is not zero** and we thus have an inconsistency. The growing magnetic field of Figure (13) is not a solution of Maxwell's equations. (This is how we play the game. Guess and try, and this time we failed.)

Equation (25) gives us a hint of what is wrong with our guess. It says that

$$\frac{\mathrm{d}\Phi_{\mathrm{E}}}{\mathrm{d}t} = -\frac{\mathrm{i}}{\varepsilon_{0}} \tag{25a}$$

thus if we have a current i and have the growing magnetic field shown in Figure (13) we must also have a changing electric flux  $\Phi_E$  through Path (a). Somewhere there must be an electric field  $\vec{E}$  to produce the changing flux  $\Phi_E$ , a field that points either up or down, passing through the circular path of Figure (13).





As a guess, we will assume that the magnetic field expands at a speed v out from the wire, when the current is turned on.

In our earlier discussion of inductance and induced voltage, we saw that a changing current creates an electric field that opposed the change. This is what gives an effective inertia to the current in an inductor. Thus when we suddenly turn on the upward directed current as shown in Figure (12), we expect that we should have a downward directed electric field as indicated in Figure (14), opposing our trying to start the current.

Initially the downward directed electric field should be inside the wire where it can act on the current carrying charges. But our growing circular magnetic field shown in Figure (13) must also have started inside the wire. Since a growing magnetic field alone is not a solution of Maxwell's equations and since there must be an associated electric field, let us propose that both the circular magnetic field of Figure (13), and the downward electric field of Figure (14) grow together as shown in Figure (15). In Figure (15), we have sketched a field structure consisting of a circular magnetic field and a downward electric field that started out at the wire and is expanding radially outward at a speed v as shown. This structure has not yet expanded out to our Path (1), so that the line integral  $\oint \vec{B} \cdot d\vec{\ell}$  is still zero and Ampere's law still requires that

$$0 = \mu_0 i + \mu_0 \varepsilon_0 \frac{d\Phi_E}{dt} \qquad Path \ 1 \ of$$
  
$$\frac{d\Phi_E}{dt} = -\frac{i}{\varepsilon_0} \qquad (26)$$

which is the same as Equation (25).

Looking at Figure (15), we see that the downward electric field gives us a negative flux  $\Phi_E$  through our path. (We chose the direction of the path so that by the right hand convention, the current i is positive.) And as the field structure expands, we have more negative flux through the path. This increasing negative flux is just what is required by Equation (26).



32-13

What happens when the field structure gets to and passes our path? The situation suddenly changes. Now we have a magnetic field at the path, so that  $\oint \vec{B} \cdot d\vec{\ell}$  is no longer zero. And now the expanding front is outside our path so that the expansion no longer contributes to  $-d\Phi_E/dt$ . The sudden appearance of  $\oint \vec{B} \cdot d\vec{\ell}$  is precisely compensated by the sudden loss of the  $d\Phi_E/dt$  due to expansion of the field structure.

The alert student, who calculates  $\oint \vec{E} \cdot d\vec{\ell}$  for some paths inside the field structure of Figure (15) will discover that we have not yet found a completely satisfactory solution to Maxwell's equations. The electric fields in close to the wire eventually die away, and only when they have gone do we get a static magnetic field given by  $\oint \vec{B} \cdot d\vec{\ell} = \mu_0 i + 0$ .

The problems associated with the electric field dying away can be avoided if we turn on the current at time t=0, and then shut it off a very short time later. In that case we should expect to see an expanding cylindrical shell of electric and magnetic fields as shown in Figure (16). The front of the shell started out when the current was turned on, and the back should start out when the current is shut off. We will guess that the front and back should both travel radially outward at a speed v as shown.



## Speed of an Electromagnetic Pulse

Let us use Figure (16), redrawn as Figure (17a), as our best guess for the structure of an electromagnetic pulse. The first step is to check that this field structure obeys Maxwell's equations. If it does, then we will see if we can solve for the speed v of the wave front.

In Figure (17a), where we have shut the current off, there is no net charge or current and all we need to consider is the expanding shell of electric and magnetic fields moving through space. We have no divergent fields, no current, and the equations for E and B become

$$\oint \vec{B} \cdot d\vec{\ell} = \mu_0 \varepsilon_0 \frac{d\Phi_E}{dt}$$
(23a)

$$\oint \vec{E} \cdot d\vec{\ell} = -\frac{d\Phi_{\rm B}}{dt}$$
(23b)

which we wrote down earlier as Maxwell's equation for empty space.



#### Figure 17a

In order to analyze the electromagnetic pulse produced by turning the current on and off, we introduce the two paths shown above. Path (1) has one side parallel to the electric field, while Path (2) has a side parallel to the magnetic field. In order to apply Maxwell's equations to the fields in Figure (17a), we will focus our attention on a small piece of the shell on the right side that is moving to the right at a speed v. For this analysis, we will use the two paths labeled Path (1) and Path (2). Path (1) has a side parallel to the electric field, and will be used for Equation 23b. Path (2) has a side parallel to the magnetic field, and will be used for Equation 23a.

#### Analysis of Path 1

In Figure (17b), we have a close up view of Path (1). The path was chosen so that only the left edge of length h was in the electric field, so that

$$\oint \vec{E} \cdot d\vec{\ell} = Eh$$
 (27)

In order to make  $\vec{E} \cdot d\vec{\ell}$  positive on this left edge, we went around Path (1) in a counterclockwise direction. By the right hand convention, any vector up through this path is positive, therefore the downward directed magnetic field is going through Path (1) in a negative direction. (We will be very careful about signs in this discussion.)





Side view showing path (1). An increasing (negative) magnetic flux flows down through Path (1).

In Figure (18), we are looking at Path (1), first at a time t (18a) where the expanding front has reached a position x as shown, then at a time  $t + \Delta t$  where the front has reached  $x + \Delta x$ . Since the front is moving at an assumed speed v, we have

$$v = \frac{\Delta x}{\Delta t}$$

At time  $t + \Delta t$ , there is additional magnetic flux through Path (1). The amount of additional magnetic flux  $\Delta \Phi_B$ is equal to the strength B of the field times the additional area ( $h\Delta x$ ). Since  $\vec{B}$  points down through Path (1), in a negative direction, the additional flux is negative and we have

$$\Delta \Phi_{\rm B} = -B(\Delta A_{\parallel}) = -B(h\Delta x) \tag{28}$$

Dividing Equation (28) through by  $\Delta t$ , and taking the limit that  $\Delta t$  goes to dt, gives

$$\frac{\Delta \Phi_{\rm B}}{\Delta t} = -Bh \frac{\Delta x}{\Delta t}$$
$$\frac{d\Phi_{\rm B}}{dt} = -Bh \frac{dx}{dt} = -Bhv \qquad (29)$$





#### Figure 18

As the front expands, there is more magnetic flux down through Path (1).

We now have a formula for  $\vec{E} \cdot d\vec{\ell}$  (Equation 27) and for  $d\Phi_E/dt$  (Equation 29) which we can substitute into Faraday's law (23b) to get

$$\oint \vec{E} \cdot d\vec{\ell} = -\frac{d\Phi_B}{dt}$$

$$Eh = -(-Bvh) = +Bhv$$

The factor of h cancels and we are left with

$$E = Bv \qquad from \\ Faraday's law \qquad (30)$$

which is a surprisingly simple relationship between the strengths of the electric and magnetic fields.

#### **Analysis of Path 2**

Path (2), shown in Figure (17c), is chosen to have one side in and parallel to the magnetic field. We have gone around clockwise so that  $\vec{B}$  and  $d\vec{\ell}$  point in the same direction. Integrating  $\vec{B}$  around the path gives

$$\oint \vec{B} \cdot d\vec{\ell} = Bh \tag{31}$$

Combining Equation 31 with Ampere's law

$$\oint \vec{\mathbf{B}} \cdot d\vec{\ell} = \mu_0 \varepsilon_0 \frac{d\Phi_E}{dt}$$

gives

$$Bh = \mu_0 \varepsilon_0 \frac{d\Phi_E}{dt}$$
(32)

To evaluate  $d\Phi_E/dt$ , we first note that for a clockwise path, the positive direction is down into the paper in Figure (17c). This is the same direction as the electric field, thus we have a positive electric flux through path (2).



**Figure 17a (repeated)** We will now turn our attention to path 2 which has one side parallel to the magnetic field.

In Figure (19), we show the expanding front at time t (19a) and at time t +  $\Delta t$  (19b). The increase in electric flux  $\Delta \Phi_E$  is (E) times the increased area (h $\Delta x$ )

$$\Delta \Phi_{\rm E} = {\rm E}({\rm h}\Delta {\rm x})$$

Dividing through by  $\Delta t$ , and taking the limit that  $\Delta t$  goes to dt, gives

$$\frac{\Delta \Phi_{\rm E}}{\Delta t} = {\rm Eh} \frac{\Delta x}{\Delta t}$$
$$\frac{d\Phi_{\rm E}}{dt} = {\rm Eh} \frac{dx}{dt} = {\rm Ehv}$$
(33)

Using Equation 33 in 32, and then cancelling h, gives

$$Bh = \mu_0 \varepsilon_0 Ehv \tag{34}$$

$$B = \mu_0 \varepsilon_0 Ev \qquad From \\ Ampere's law$$

which is another simple relationship between E and B.



**Figure 17c** An increasing (negative) electric flux flows down through Path (2).

If we divide Equation (30) Bv = E, by Equation (34)  $B = \mu_0 \varepsilon_0 Ev$ , both E and B cancel giving

$$\frac{Bv}{B} = \frac{E}{\mu_0 \varepsilon_0 E v}$$
$$v^2 = \frac{1}{\mu_0 \varepsilon_0}$$
$$speed of light!!!$$

Thus the electromagnetic pulse of Figures (16) and (17) expands outward at the speed  $1/\sqrt{\mu_0\epsilon_0}$  which we have seen is  $3 \times 10^8$  meters per second. Maxwell recognized that this was the speed of light and recognized that the electromagnetic pulse must be closely related to light itself.

Using 
$$v = 1/\sqrt{\mu_0 \varepsilon_0} = c$$
 in Equation (34) we get  

$$B = \frac{E}{c}$$
(36)

as the relative strength of the electric and magnetic fields in an electromagnetic pulse, or as we shall see, any light wave. If we had used a reasonable set of units where c = 1 (like feet and nanoseconds), then E and B would have equal strengths in a light wave.



Figure 19

As the front expands, there is more electric flux down through Path (2).

#### **Exercise 4**

Construct paths like (1) and (2) of Figure (17), but which include the back side, rather than the front side, of the electromagnetic pulse. Repeat the kind of steps used to derive Equation (35) to show that the back of the pulse also travels outward at a speed  $v = 1/\sqrt{\mu_0 \epsilon_0}$ . As a result the pulse maintains its thickness as it expands out through space.

#### Exercise 5

(35)

After a class in which we discussed the electromagnetic pulse shown in Figure (20a), a student said she thought that the electric field would get ahead of the magnetic field as shown in Figure (20b). Use Maxwell's equations to show that this does not happen.



#### Figure 20a

The radiated electromagnetic pulse we saw in Figures (16) and (17).



#### Figure 20b

The student guessed that the electric fields would get out ahead of the magnetic field. Use Path (1) to show that this does not happen.

## **ELECTROMAGNETIC WAVES**

The single electromagnetic pulse shown in Figure (17) is an example of an electromagnetic wave. We usually think of a wave as some kind of oscillating sinusoidal thing, but as we saw in our discussion of waves on a Slinky in Chapter 1, the simplest form of a wave is a single pulse like that shown in Figure (21). The basic feature of the Slinky wave pulse was that it maintained its shape while it moved down the Slinky at the wave speed v. Now we see that the electromagnetic pulse maintains its structure of  $\vec{E}$  and  $\vec{B}$  fields while it moves at a speed v = c through space.

We made a more or less sinusoidal wave on the Slinky by shaking one end up and down to produce a series of alternate up and down pulses that traveled together down the Slinky. Similarly, if we use an alternating current in the wire of Figure (17), we will get a series of electromagnetic pulses that travel out from the wire. This series of pulses will more closely resemble what we usually think of as an electromagnetic wave.

Figure (22a) is a graph of a rather jerky alternating current where we turn on an upward directed current of magnitude  $i_0$ , then shut off the current for a while, then turn on a downward directed current  $i_0$ , etc. This series of current pulse produces the series of electromagnetic pulses shown in Figure (22b). Far out from the wire where we can neglect the curvature of the magnetic field, we see a series of pulses shown in the close-up view, Figure (23a). This series of flat or non-curved pulses is called a *plane wave* of electromagnetic radiation.



Figure 21 Slinky wave pulse.

If we used a sinusoidally oscillating current in the wire of Figure (22), then the series of electromagnetic pulses would blend together to form the sinusoidally varying electric and magnetic fields structure shown in Figure (23b). This is the wave structure one usually associates with an electromagnetic wave.

When you think of an electromagnetic wave, picture the fields shown in Figure (23), moving more or less as a rigid object past you at a speed c. The distance  $\lambda$  between crests is called the *wavelength* of the wave. The time T it takes one wavelength or cycle to pass you is

$$T\frac{\text{second}}{\text{cycle}} = \frac{\lambda \frac{\text{meter}}{\text{cycle}}}{c \frac{\text{meter}}{\text{second}}} = \frac{\lambda \text{ second}}{c \text{ cycle}}$$
(37)

T is called the *period* of the wave. The *frequency* of the wave, the number of wavelengths or full cycles of the wave that pass you per second is

$$f\frac{\text{cycle}}{\text{second}} = \frac{c \frac{\text{meter}}{\text{second}}}{\lambda \frac{\text{meter}}{\text{cycle}}} = \frac{\lambda \text{ cycle}}{c \text{ second}}$$
(38)

In Equations (37) and (38) we gave  $\lambda$  the dimensions meters/cycle, T of seconds/cycle and f of cycles/second so that we can use the dimensions to remember the





b) Resulting electric and magnetic fields

## **Figure 22** *Fields produced by a series of current pulses.*

formulas  $T = \lambda/c$ ,  $f = c/\lambda$ . (It is now common to use "hertz" or "Hz" for the dimensions of frequency. This is a classic example of ruining simple dimensional analysis by using people's names.) Finally, the angular frequency  $\omega$  radians per second is defined as

$$\omega \frac{\text{radians}}{\text{sec ond}} = 2\pi \frac{\text{radians}}{\text{cycle}} \times f \frac{\text{cycles}}{\text{second}}$$
$$= 2\pi f \frac{\text{radians}}{\text{second}}$$
(39)

You can remember where the  $2\pi$  goes by giving it the dimensions  $2\pi$  radians/cycle. (Think of a full circle or full cycle as having  $2\pi$  radians.) We will indiscriminately use the word *frequency* to describe either f cycles/second or  $\omega$  radians/second, whichever is more appropriate. If, however, we say that something has a frequency of so many hertz, as in 60 Hz, we will always mean cycles/second.



**Figure 23** *Structure of electric and magnetic fields in light and radio waves.* 

## ELECTROMAGNETIC SPECTRUM

We have seen by direct calculation that the electromagnetic pulse of Figure (17), and the series of pulses in Figure (22) are a solution of Maxwell's equations. It is not much of an extension of our work to show that the sinusoidal wave structure of Figure (23b) is also a solution. The fact that all of these structures move at a speed  $c = 1/\sqrt{\mu_0 \varepsilon_0} = 3 \times 10^8$  m/s is what suggested to Maxwell that these electromagnetic waves were light, that he had discovered the theory of light.

But there is nothing in Maxwell's equations that restricts our sinusoidal solution in Figure (23b) to certain values or ranges of frequency or wavelength. One hundred years before Maxwell it was known from interference experiments (which we will discuss in the next chapter) that light had a wave nature and that the wavelengths of light ranged from about  $6 \times 10^{-5}$  cm in the red part of the spectrum down to  $4 \times 10^{-5}$  cm in the blue part. With the discovery of Maxwell's theory of light, it became clear that there must be a complete spectrum of electromagnetic radiation, from very long down to very short wavelengths, and that visible light was just a tiny piece of this spectrum.

More importantly, Maxwell's theory provided the clue as to how you might be able to create electromagnetic waves at other frequencies. We have seen that an oscillating current in a wire produces an electromagnetic wave whose frequency is the same as that of the current. If, for example, the frequency of the current is 1030 kc (1030 kilocycles) =  $1.03 \times 10^6$  cycles/sec, then the electromagnetic wave produced should have a wavelength

$$\lambda \frac{\text{meters}}{\text{cycle}} = \frac{c \frac{\text{meters}}{\text{second}}}{f \frac{\text{cycles}}{\text{sec ond}}} = \frac{3 \times 10^8 \text{m/s}}{103 \times 10^6 \text{c/s}}$$
$$= 297 \text{ meters}$$

Such waves were discovered within 10 years of Maxwell's theory, and were called radio waves. The frequency 1030 kc is the frequency of radio station WBZ in Boston, Mass.

## Components of the Electromagnetic Spectrum

Figure (24) shows the complete electromagnetic spectrum as we know it today. We have labeled various components that may be familiar to the reader. These components, and the corresponding range of wavelengths are as follows:

Radio Waves	10 <sup>6</sup> m to .05 mm
AM Band	500 m to 190 m
Short Wave	60 m to 15 m
TV VHF Band	10 m to 1 m
TV UHF Band	1 m to 10 cm
Microwaves	10 cm to .05 mm
Infrared Light	.05 mm to $6 \times 10^{-5}$ cm
Visible Light	$6 \times 10^{-5}$ to $4 \times 10^{-5}$ cm
Ultraviolet Light	$4 \times 10^{-5}$ cm to $10^{-6}$ cm
X Rays	10 $^{-6}$ cm to 10 $^{-9}$ cm
γ Rays	$10^{-9}$ cm and shorter



#### Figure 24

The electromagnetic spectrum extends from long wavelength radio waves down to short wavelength X rays and gamma rays. The visible part of the spectrum is indicated by the small box.

In each of these ranges, the most efficient way to emit or detect the radiation is to use antennas whose size is comparable to the wavelength of the radiation. For radio waves the antennas are generally some kind of a structure made from wire. In the infrared and the visible region, radiation is generally emitted by molecules and atoms. The short wavelength x rays and  $\gamma$ rays generally come from atomic nuclei or subatomic particles.

The longest wavelength radio waves that have been studied are the so-called "whistlers", radio waves with an audio frequency, that are produced by lightening bolts and reflected back and forth around the earth by charged particles trapped in the earth's magnetic field. On a shorter scale of distance are the long wavelength radio waves which penetrate the ocean and are used for communications with submarines. The radio station in Cutler, Maine, shown in Figure (25), has twenty-six towers over 1000 feet tall to support the antenna to produce such waves. This station, operated by the United States Navy, is the world's most powerful.

As we go to shorter wavelengths and smaller antennas, we get to the broadcast band, short wave radio, then to the VHF and UHF television frequencies. (FM radio is tucked into the VHF band next to Channel 6). The wavelengths for VHF television are of the order of meters, while those for UHF are of the order of a foot. Those with separate VHF and UHF television antennas will be familiar with the fact that the UHF antenna, which detects the shorter wavelengths, is smaller in size.

Adjusting the rabbit ears antenna on a television set provides practical experience with the problems of detecting an electromagnetic wave. As the TV signal strikes the antenna, the electric field in the wave acts on the electrons in the TV antenna wire. If the wire is parallel to the electric field, the electrons are pushed along in the wire producing a voltage that is detected by the television set. If the wire is perpendicular, the electrons will not be pushed up and down and no voltage will be produced.

The length of the wire is also important. If the antenna were one half wavelength, then the electric field at one end would be pushing in the opposite direction from the field at the other end, the integral  $\oint \vec{E} \cdot \vec{dl}$  down the antenna would be zero, and you would get no net voltage or signal. You want the antenna long enough to get a big voltage, but not so long that the electric field in one part of the antenna works against the field in another part. One quarter wavelength is generally the optimum antenna length.



#### Figure 25

The worlds largest radio station at Cutler, Maine. This structure, with 75 miles of antenna wire and 26 towers over 1000 ft high, generates long wavelength low frequency, radio waves for communications with submarines.

The microwave region, now familiar from microwave communications and particularly microwave ovens, lies between the television frequencies and infrared radiation. The fact that you heat food in a microwave oven emphasizes the fact that electromagnetic radiation carries energy. One can derive that the energy density in an electromagnetic wave is given by the formula

energy density in an electromagnetic wave 
$$\left. \right\rangle = \frac{\varepsilon_0 E^2}{2} + \frac{B^2}{2\mu_0} \quad (37)$$

We have already seen the first term  $\epsilon_0 E^2/2$ , when we calculated the energy stored in a capacitor (see Equation 27-36 on page 27-19). If we had calculated the energy to start a current in an inductor, we would have gotten the formula  $B^2/2\mu_0$  for the energy density in that device. Equation (37) tells us the amount of energy is associated with electric and magnetic fields whenever we find them.

## **Blackbody Radiation**

Atoms and molecules emit radiation in the infrared, visible and ultraviolet part of the spectrum. One of the main sources of radiation in this part of the spectrum is the so-called *blackbody* radiation emitted by objects due to the thermal motion of their atoms and molecules.

If you heat an iron poker in a fire, the poker first gets warm, then begins to glow a dull red, then a bright red or even, orange. At higher temperatures the poker becomes white, like the filaments in an electric light bulb. At still higher temperatures, if the poker did not melt, it would become bluish. The name *blackbody radiation* is related to the fact that an initially cold, black object emits these colors of light when heated.

There is a well studied relationship between the temperature of an object and the predominant frequency of the blackbody radiation it emits. Basically, the higher the temperature, the higher the frequency. Astronomers use this relationship to determine the temperature of stars from their color. The infrared stars are quite cool, our yellow sun has about the same temperature as the yellow filament in an incandescent lamp, and the blue stars are the hottest. All objects emit blackbody radiation. You, yourself, are like a small star emitting infrared radiation at a wavelength corresponding to a temperature of 300K. In an infrared photograph taken at night, you would show up distinctly due to this radiation. Infrared photographs are now taken of houses at night to show up hot spots and heat leaks in the house.

Perhaps the most famous example of blackbody radiation is the 3K cosmic background radiation which is the remnant of the big bang which created the universe. We will say much more about this radiation in Chapter 34.

## UV, X Rays, and Gamma Rays

When we get to wavelengths shorter than the visible spectrum, and even in the visible spectrum, we begin to run into problems with Maxwell's theory of light. These problems were first clearly displayed by Max Planck who in 1900 developed a theory that explained the blackbody spectrum of radiation. The problem with Planck's theory of blackbody radiation is that it could not be derived from Maxwell's theory of light and Newtonian mechanics. His theory involved arbitrary assumptions that would not be understood for another 23 years, until after the development of quantum mechanics.

Despite the failure of Newton's and Maxwell's theories to explain all the details, the electromagnetic spectrum continues right on up into the shorter wavelengths of ultraviolet (UV) light, then to x rays and finally to  $\gamma$  rays. Ultraviolet light is most familiar from the effect it has on us, causing tanning, sunburns, and skin cancer depending on the intensity and duration of the dose. The ozone layer in the upper atmosphere, as long as it lasts, is important because it filters out much of the ultraviolet light emitted by the sun.

X rays are famous for their ability to penetrate flesh and produce photographs of bones. These rays are usually emitted by the tightly bound electrons on the inside of large atoms, and also by nuclear reactions. The highest frequency radiation,  $\gamma$  rays, are emitted by the smallest objects—nuclei and elementary particles.

## POLARIZATION

One of the immediate tests of our picture of a light or radio wave, shown in Figure (23), is the phenomena of *polarization*. We mentioned that the reason that you had to adjust the angle of the wires on a rabbit ears antenna was that the electric field of the television signal had to have a significant component parallel to the wires in order to push the electrons up and down the wire. Or, in the terminology of the last few chapters, we needed the parallel component of  $\vec{E}$  so that the voltage  $V = \oint \vec{E} \cdot \vec{d1}$  would be large enough to be detected by the television circuitry. (In this case, the line integral  $\oint \vec{E} \cdot \vec{d1}$  is along the antenna wire.) Polarization is a phenomena that results from the fact that the electric field  $\vec{E}$  in an electromagnetic wave can have various orientations as the wave moves through space.

Although we have derived the structure of an electromagnetic wave for the specific case of a wave produced by an alternating current in a long, straight wire, some of the general features of electromagnetic waves are clearly present in our solution. The general features that are present in all electromagnetic waves are:

- 1) All electromagnetic waves are a structure consisting of an electric field  $\vec{E}$  and a magnetic field  $\vec{B}$ .
- 2)  $\vec{E}$  and  $\vec{B}$  are at right angles to each other as shown in Figure (23).
- 3) The wave travels in a direction perpendicular to the plane of  $\vec{E}$  and  $\vec{B}$ .
- 4) The speed of the wave is  $c = 3 \times 10^8 m/s$ .
- 5) The relative strengths of  $\vec{E}$  and  $\vec{B}$  are given by Equation (36) as B = E/c.

Even with these restrictions, and even if we consider only flat or plane electromagnetic waves, there are still various possible orientations of the electric field as shown in Figure (26). In Figure (26a) we see a plane wave with a vertical electrical field. This would be called a vertically polarized wave. In Figure (26b), where the electric field is horizontal, we have a horizontally polarized wave. By convention we say that the direction of polarization is the direction of the electric field in an electromagnetic wave. Because  $\vec{E}$  must lie in the plane perpendicular to the direction of motion of an electromagnetic wave,  $\vec{E}$  has only two independent components, which we can call the vertical and horizontal polarizations, or the x and y polarizations as shown in Figures (27a) and (27b) respectively. If we happen to encounter an electromagnetic wave where  $\vec{E}$  is neither vertical or horizontal, but at some angle  $\theta$ , we can decompose  $\vec{E}$  into its x and y components as shown in (27c). Thus we can consider a wave polarized at an arbitrary angle  $\theta$  as a mixture of the two independent polarizations.



#### b) Horizontally polarized electromagnetic wave.

#### **Figure 26** *Two possible polarizations of an electromagnetic wave.*

Figure 27

We define the direction

electromagnetic wave as

of polarization of an

the direction of the

electric field.

€ ↑

## \_\_\_\_**↓**Ē

b) Horizontal Polarization



c) Mixture

## Polarizers

A polarizer is a device that lets only one of the two possible polarizations of an electromagnetic wave pass through. If we are working with microwaves whose wavelength is of the order of a few centimeters, a frame strung with parallel copper wires, as seen in Figure (28), makes an excellent polarizer. If a vertically polarized wave strikes this vertical array of wires, the electric field  $\vec{E}$  in the wave will be parallel to the wires. This parallel  $\vec{E}$  field will cause electrons to move up and down in the wires, taking energy out of the incident wave. As a result the vertically polarized wave cannot get through. (One can observe that the wave is actually reflected by the parallel wires.) If you then rotate the wires 90°, so that the  $\vec{E}$  field in the wave is perpendicular to the wires, the electric field can no longer move electrons along the wires and the wires have no effect. The wave passes through without attenuation.

If you do not happen to know the direction of polarization of the microwave, put the polarizer in the beam and rotate it. For one orientation the microwave beam will be completely blocked. Rotate the polarizer by  $90^{\circ}$  and you will get a maximum transmission.



## Figure 28

Microwave polarizer, made from an array of copper wires. The microwave transmitter is seen on the other side of the wires, the detector is on this side. When the wires are parallel to the transmitted electric field, no signal is detected. Rotate the wires 90 degrees, and the full signal is detected.

#### **Light Polarizers**

We can picture light from the sun as a mixture of light waves with randomly oriented polarizations. (The  $\vec{E}$ fields are, of course, always in the plane perpendicular to the direction of motion of the light wave. Only the angle in that plane is random.) A polarizer made of an array of copper wires like that shown in Figure (28), will not work for light because the wavelength of light is so short ( $\lambda \approx 5 \times 10^{-5}$  cm) that the light passes right between the wires. For such a polarizer to be effective, the spacing between the wires would have to be of the order of a wavelength of light or less.

A polarizer for light can be constructed by imbedding long-chain molecules in a flexible plastic sheet, and then stretching the sheet so that the molecules are aligned parallel to each other. The molecules act like the wires in our copper wire array, but have a spacing of the order of the wavelength of light. As a result the molecules block light waves whose electric field is parallel to them, while allowing waves with a perpendicular electric field to pass. (The commercial name for such a sheet of plastic is *Polaroid*.) Since light from the sun or from standard electric light bulbs consists of many randomly polarized waves, a single sheet of Polaroid removes half of the waves no matter how we orient the Polaroid (as long as the sheet of Polaroid is perpendicular to the direction of motion of the light beam). But once the light has gone through one sheet of Polaroid, all the surviving light waves have the same polarization. If we place a second sheet of Polaroid over the first, all the light will be absorbed if the long molecules in the second sheet are perpendicular to the long molecules in the first sheet. If the long molecules in the second sheet are parallel to those in the first, most of the waves that make it through the first, make it through the second also. This effect is seen clearly in Figure (29).



#### Figure 29

Light polarizers. Two sheets of polaroid are placed on top of a drawing. On the left, the axes of the sheets are parallel, so that nearly half the light passes through. On the right, the axes are perpendicular, so that no light passes through. (Photo from Halliday & Resnick)

## **Magnetic Field Detector**

So far, our discussion of electromagnetic radiation has focused primarily on detecting the *electric field* in the wave. The rabbit ear antenna wire had to be partially parallel to the electric field so that  $\oint \vec{E} \cdot d\vec{l}$  and therefore the voltage on the antenna would not be zero. In our discussion of polarization, we aligned the parallel array of wires or molecules parallel to the electric field when we wanted the radiation to be reflected or absorbed.

It is also fairly easy to detect the **magnetic field** in a radio wave by using one of our  $\oint \vec{E} \cdot d\vec{\ell}$  meters to detect a changing magnetic flux (an application of Faraday's law). This is the principle behind the radio direction finders featured in a few World War II spy pictures.



**Figure 30a** Car with radio direction finder loop antenna mounted on top.

#### Figure 31

In a January 1998 National Geographic article on Amelia Earhardt, there appeared a picture of a vintage Electra airplane similar to the one flown by Earhardt on her last trip in 1938. On the top of the plane, you can see the kind of radio direction finder we have been discussing. (The plane is being flown by Linda Finch.) In a typical scene we see a car with a metal loop mounted on top as shown in Figure (30a). It is chasing another car with a hidden transmitter, or looking for a clandestine enemy transmitter.

If the transmitter is a radio antenna with a vertical transmitting wire as shown in Figure (30b), the magnetic field of the radiated wave will be concentric circles as shown. Objects on the ground, the ground itself, and nearby buildings and hills can distort this picture, but for now we will neglect the distortions.



**Figure 30b** *Car driving toward radio transmitter.* 



32-27

In Figures (32a) and (32b), we show the magnetic field of the radio wave as it passes the detector loop mounted on the car. A voltmeter is attached to the loop as shown in Figure (33). In (32a), the plane of the loop is parallel to  $\vec{B}$ , the magnetic flux  $\Phi_B$  through the loop is zero, and Faraday's law gives

$$V = \oint \vec{E} \cdot d\vec{\ell} = \frac{d\Phi_{\rm B}}{dt} = 0$$

In this orientation there is no voltage reading on the voltmeter attached to the loop.

In the orientation of Figure (32b), the magnetic field passes through the loop and we get a maximum amount of magnetic flux  $\Phi_B$ . As the radio wave passes by the loop, this flux alternates signs at the frequency of the wave, therefore the rate of change of flux  $d\Phi_B/dt$  is at a maximum. In this orientation we get a maximum voltmeter reading.

The most sensitive way to use this radio direction finder is to get a zero or "null" reading on the voltmeter. Only when the loop is oriented as in Figure (32a), with its plane perpendicular to the direction of motion of the radio wave, will we get a null reading. At any other orientation some magnetic flux will pass through the loop and we get some voltage.

Spy pictures, set in more modern times, do not show antenna loops like that in Figure (30) because modern radio direction finders use so-called "ferrite" antennas that detect the electric field in the radio wave. We get a voltage on a ferrite antenna when the electric field in the radio wave has a component along the ferrite rod, just as it needed a component along the wires of a rabbit ears antenna. Again these direction finders are most accurate when detecting a null or zero voltage. This occurs only when the rod is parallel to the direction of motion of the radio wave, i.e. points toward the station. (This effect is very obvious in a small portable radio. You will notice that the reception disappears and you get a null detection, for some orientations of the radio.)



Electric

 b) Loop oriented so that magnetic flux goes through it

#### Figure 32

Magnetic

Electromagnetic field impinging upon a loop antenna. In (a), the magnetic field is parallel to the plane of the loop, and therefore no magnetic flux goes through the loop. In (b), the magnetic flux goes through the loop. As the wave passes by, the amount of flux changes, inducing a voltage in the loop antenna.



## Figure 33

We can think of a wire loop connected to a voltmeter as an  $\oint \vec{E} d\ell$  meter. Any changing magnetic flux through the loop induces a voltage around the loop. This voltage is read by the voltmeter.

## **RADIATED ELECTRIC FIELDS**

One of the best computer simulations of physical phenomena is the series of short films about the electric fields produced by moving and accelerated charges. We will describe a few of the frames from these films, but nothing replaces watching them.

Two basic ideas underlie these films. One is Gauss' law which requires that *electric field lines not break*, do not end, in empty space. The other is that disturbances on an *electric field line travel outward at the speed of light*. No disturbance, no change in the electric field structure, can travel faster than the speed of light without violating causality. (You could get answers to questions that have not yet been asked.)

As an introduction to the computer simulations of radiation, let us see how a simple application of these two basic ideas leads to the picture of the electromagnetic pulse shown back in Figure (16). In Figure (34a) we show the electric field of a stationary, positively charged rod. The electric field lines go radially outward to infinity. (It's a long rod, and it has been at rest for a long time.)

At time t = 0 we start moving the entire rod upward at a speed v. By Gauss' law the electric field lines must stay attached to the charges Q in the rod, so that the ends of the electric field lines have to start moving up with the rod.

No information about our moving the rod can travel outward from the rod faster than the speed of light. If the time is now t > 0, then beyond a distance ct, the electric field lines must still be radially outward as in Figure (34b). To keep the field lines radial beyond r = ct, and keep them attached to the charges +Q in the rod, there must be some kind of expanding kink in the lines as indicated. At time  $t = t_1$ , we stop moving the positively charged rod. The information that the charged rod has stopped moving cannot travel faster than the speed of light, thus the displaced radial field next to the rod cannot be any farther out than a distance  $c(t-t_1)$  as shown in Figure (34c). The effect of starting, then stopping the positive rod is an outward traveling kink in the electric field lines. It is as if we had ropes attached to the positive rod, and jerking the rod produced an outward traveling kink or wave on the ropes.

In Figure (34d), we have added in a stationary negatively charged rod and the inward directed electric field produced by that rod. The charge density on the negative rod is opposite that of the positive rod, so that there is no net charge on the two rods. When we combine these rods, all we have left is a positive upward directed current during the time interval t=0 to  $t=t_1$ . We have a short current pulse, and the electric field produced by the current pulse must be the vector sum of the electric fields of the two rods.

In Figure (34e), we add up the two electric fields. In the region r > ct beyond the kink, the positive and negative fields must cancel exactly. In the region  $r < (t - t_1)$  we should also have nearly complete cancellation. Thus all we are left with are the fields  $E_+$  and  $E_-$  inside the kink as shown in Figure (34f). Since electric field lines cannot end in empty space,  $E_+$  and  $E_-$  must add up to produce the downward directed  $E_{net}$  shown in Figure (34g). Note that this downward directed electric field pulse was produced by an upward directed current pulse. As we have seen before, this induced electric field opposes the change in current.

In Figure (34h) we added the expanding magnetic field pulse that should be associated with the current pulse. What we see is an expanding electromagnetic pulse that has the structure shown in Figure (16). Simple arguments based on Gauss' law and causality gave us most of the results we worked so hard to get earlier. What we did get earlier, however, when we applied Ampere's and Faraday's law to this field structure, was the explicit prediction that the pulse expands at the speed  $1/\sqrt{\mu_0}\varepsilon_0 = 3 \times 10^8$  m/s.



a) Electric field of a stationary, positively charged rod.



b) At time t = 0, we start moving the entire rod upward at a speed v. The ends of the field lines must stay attached to the charges in the rod.



c) We stop moving the rod at time  $t = t_1$ . No information about our having moved the rod can travel out faster than the speed of light.



d) We have added in the electric field of a line of stationary negative charges. As a result, the net charge on the rod is zero and we have only a current pulse that lasted from t = 0 to  $t = t_1$ .



e) When we add up the electric fields of the positive and negative rods, the fields cancel everywhere except at the outward going pulse.



f) At the pulse, the vector sum of  $\vec{E}_+$  and  $\vec{E}_-$  is a downward directed field  $\vec{E}_{net}$  as shown.



g) Thus a short upward directed current pulse produces a downward directed electric field that travels outward from the wire at a speed c.



h) Add in the magnetic field of the current pulse, and we have the electromagnetic wave structure seen in Figure (16).

#### Figure 34

+

Using the fact that electric field lines cannot break in empty space (Gauss' Law), and the idea that kinks in the field lines travel at the speed of light, we can guess the structure of an electromagnetic pulse.
# **Field of a Point Charge**

The computer simulations show the electric field of a point charge under varying situations. In the first, we see the electric field of a point charge at rest, as shown in Figure (35a). Then we see a charge moving at constant velocity  $\vec{v}$ . As the speed of the charge approaches c, the electric field scrunches up as shown in Figure (35b).

The next film segment shows what happens when we have a moving charge that stops. If the charge stopped at time t = 0, then at a distance r = ct or greater, we must have the electric field of a moving charge, because no information that the charge has stopped can reach beyond this distance. In close we have the electric field of a static charge. The expanding kink that connects the two regions is the electromagnetic wave. The result is shown in Figure (35c). The final film segment shows the electric field of an oscillating charge. Figure (36) shows one frame of the film. This still picture does a serious injustice to the animated film. There is no substitute, or words to explain, what you see and feel when you watch this film.



**Figure 35a** *Electric field of a stationary charge.* 



#### Figure 35b

Electric field of a moving charge. If the charge has been moving at constant speed for a long time, the field is radial, but squeezed up at the top and bottom.



#### Figure 35c

Field of a charge that stopped. Assume that the charge stopped t seconds ago. Inside a circle of radius ct, we have the field of a stationary charge. Outside, where there is no information that the charge has stopped, we still have the field of a moving charge. The kink that connects the two fields is the electromagnetic radiation.



Figure 35c (enlarged)

Electric field of a charge that stopped. The dotted lines show the field structure we would have seen had the charge not stopped.

**Figure 36** *Electric field of an oscillating charge.* 

#### **Exercise 6**

Assume that we have a supply of ping pong balls and cardboard tubes shown in Figures (37). By looking at the fields outside these objects decide what could be inside producing the fields. Explicitly do the following for each case.

i) Write down the Maxwell equation which you used to decide what is inside the ball or tube, and explain how you used the equation.

ii) If more than one kind of source could produce the field shown, describe both (or all) sources and show the appropriate Maxwell equations.

iii) If the field is impossible, explain why, using a Maxwell equation to back up your explanation.

In each case, we have indicated whether the source is in a ball or tube. Magnetic fields are dashed lines, electric fields are solid lines, and the balls and tubes are surrounded by empty space.



**Figure 37a** Electric field emerging from ping pong ball.



**Figure 37b** *Magnetic field emerging from ping pong ball.* 



**Figure 37c** *Electric field emerging above ping pong ball.* 



**Figure 37d** *Electric field around tube.* 





For this example, explain what is happening to the fields, what is in the tube, and what happened inside.



**Figure 37g** *There is only ONE object inside this tube.* 

What is it? What is it doing?





# Chapter 33 Light Waves



Ripples produced by rain drops. (Bill Jack Rodgers, Los Alamos Scientific Laboratory)

In the examples of wave motion we studied back in Chapter 15, like waves on a rope and sound in a gas, we could picture the wave motion as a consequence of the mechanical behavior particles in the rope or molecules in the gas. We used Newton's laws to predict the speed of a rope wave and could have done the same for a sound. When we discuss light waves, we go beyond the Newtonian behavior. Waves on a rope, on water or in a gas are mechanical undulations of an explicit medium. Light waves travel through empty space; there is nothing to undulate, nothing to which we can apply Newton's laws. Yet, in many ways, the behavior of light waves, water waves, sound waves, and even the waves of quantum theory, are remarkably similar. There are general rules of wave motion that transcend the nature of the medium or type of wave. One is the principle of superposition that we used extensively in Chapter 15. It is the idea that as waves move through each other, they produce an overall wave whose amplitude is the sum of the amplitudes of the individual waves. The other is a concept we will use extensively in this chapter called the **Huygens principle**, named after its discoverer Christian Huygens, a contemporary of Isaac Newton.

We will see that a straightforward application of the principle of superposition and the Huygens principle allows us to make detailed predictions that even can be used as a test of the wave nature of the phenomena we are studying.

# SUPERPOSITION OF CIRCULAR WAVE PATTERNS

When we studied the interaction of waves on a rope, it was a relatively simple process of adding up the individual waves to see what the resultant wave would be. For example, in Figure (15-6) reproduced here, we see that when a crest and a trough run into each other, for an instant they add up to produce a flat rope. At this instant the crest and the trough cancel each other. In contrast two crests add to produce a big crest, and two troughs add to produce a deeper trough.



When we extend our study of wave motion to two and three dimensions, the principle of superposition works the same way, but now we have to add patterns rather than just heights along a line. If, for example, we are studying wave motion on the surface of water, and two wave patterns move through each other, the resulting wave is the sum of the heights of the individual waves at every point on the surface. We do the same addition as we did for one dimensional waves, but at many more points.

A relatively simple, but important example of the superposition of wave patterns is the pattern we get when concentric circular waves from two nearby sources run into each other. The pattern is easy to set up in a ripple tank using two oscillating plungers.

Figure (1a) shows the circular wave pattern produced by a single oscillating plunger. From this picture we can easily see the circular waves emerging from the plunger. The only difficulty is distinguishing crests from troughs. We will handle this by using a solid line to represent the crest of a wave and a dashed line for a trough, as illustrated in Figure (1b).

Figure 15-6

When a crest meets a trough, there is a short time when the waves cancel.



Circular wave pattern produced in a ripple tank by a plunger. The pattern consists of alternate crests and troughs. To diagram the circular wave pattern, we will use solid lines for crests and dashed lines for troughs.





In Figure (2a), we see the wave pattern produced by two plungers oscillating side by side. Each plunger sends out a circular set of waves like that seen in Figure (1). When the two sets of circular waves cross each other, we get cancellation where crests from one set meet troughs from the other set (where a solid line from one set of circles meets a dashed line from the other set of circles in Figure (2b). This cancellation occurs along lines called *lines of nodes* which are clearly seen in Figure (2a). Between the lines of nodes we get beams of waves. In each beam, crests from one plunger meet crests from the other producing a higher crest. And troughs from one set meet troughs from the other producing deeper troughs. In our drawing of circles, Figures (2b) and (2c), we get beams of waves along the lines where solid circles cross solid circles and dashed circles cross dashed circles.





## a)

#### Figure 2a,b

Ripple tank photograph of an interference pattern. When two sets of circular waves move through each other, there are lines along which crests from one set always meet troughs from the other set. These are called lines of nodes. Between the lines of nodes, we get beams of waves. The resulting pattern is called an interference pattern.



#### Figure 2c

We get beams of waves where crests meet crests and troughs meet troughs. The lines of nodes are where crests meet troughs and the waves cancel.

# **HUYGENS PRINCIPLE**

When sunlight streams in through an open kitchen door, we see a distinct shadow on the floor. The shadow can be explained by assuming that the light beams travel in straight lines from the sun through the doorway. The whole subject of geometrical optics and lens design is based on the assumption that light travels in straight lines (except at the interface of two media of different indices of refraction).

In Figure (3) we see what happens when a wave impinges upon a slit whose width is comparable to the wavelength of the waves. Instead of there being a shadow of the slit, we see that the emerging wave comes out in all directions. The wave pattern on the right side of the slit is essentially identical to the wave pattern produced by the oscillating plunger in Figure (1a). We can explain Figure (3) by saying that the small piece of wave front that gets through the slit acts as a source of waves in much the same way that the oscillating plunger acted as a source of waves.

Christian Huygens noted this phenomena and from it developed a general principle of wave motion. His idea was that as a wave pattern evolved, each point of a wave front acts as the source of a new circular or spherical wave. To see how this principle can be applied,



#### Figure 3

A wave emerging from a narrow slit spreads out in all directions, just as if the wave in the slit were a plunger.

consider the relatively smooth wave front shown in Figure (4). To predict the position of the wave front a short time later, we treat each point on the front of the wave as a source of circular waves. We can see the effect by drawing a series of circles at closely spaced points along the wave. The circular waves add up to produce a new wave front farther out. While you can use the same construction to figure out what is happening throughout the wave, it is much easier to see what is happening at the front.

# Exercise 1

At some instant of time, the front of a wave has a sharp, right angle corner. Use Huygens principle to find the shape of the wave front at some later instant of time. (Draw a right angle corner and use the kind of construction shown in Figure (4).)



#### Figure 4

Huygen's construction. The future position of a wavefront can accurately be predicted by assuming that each point on the wavefront is a source of a new wave. By using the construction of Figure (4) to predict the future shape of a wave front, we see that if we use a slit to block all but a small section of the wave front, as illustrated in Figure (5), then the remaining piece of wave front will act as a source of circular waves emerging from the other side. This is what we saw in Figure (3). Thus the Huygens construction allows us to see not only how a smooth wave travels forward intact, but also why circular waves emerge from a narrow slit as we saw in Figure (3).

The Huygens construction also provides a picture of what happens as waves go through progressively wider slits. If the slit is wider than a wavelength then we have more sources in the slit and the waves from the sources begin to interfere with each other. In Figures (6, 7, 8)we see the wave patterns for increasingly wide slits and the corresponding Huygens constructions. For the wider slits, more of the wave goes through the center intact, but there is always a circular wave coming out at the edges. For the slit of Figure (8), the circular waves at the edges are relatively unimportant, and the edges of the slit cast a shadow. This is beginning to resemble our example of sunlight coming through the kitchen doorway. The name *diffraction* is used to describe the spreading of the waves that we see at the edges of the slits in Figures (5) through (8).





**Figure 5** The small piece of wave in a narrow slit acting as a single point source.



**Figure 6** When the slit is about 2 wavelengths wide, the wave in the slit acts as 2 point sources.



#### Figure 7

As the slit is widened, more of the wave comes through intact. In the center we are beginning to get a beam of waves, yet at the edges, the wavefront continues to act as a source of circular waves.



#### Figure 8

When the slit is wide compared to a wavelength, we get a distinct beam of waves. Yet no matter how wide the slit, there are still circular waves at the edges.

# TWO SLIT INTERFERENCE PATTERN

If a single narrow slit can produce the same wave pattern as an oscillating plunger, as we saw in Figure (3), then we should expect that two slits next to each other should produce an interference pattern similar to the one produced by two oscillating plungers seen in Figure (2). That this is indeed correct is demonstrated in Figure (9). On the left we have repeated the wave pattern of 2 plungers. On the right we have a wave impinging upon two narrow slits. We see that both have the same structure of lines of nodes, with beam of waves coming out between the lines of nodes. Because the patterns are the same, we can use the same analysis for both situations. Sending a wave through two slits and observing the resulting wave pattern is a convenient way to analyze various kinds of wave motion. But in most cases we do not see the full interference pattern as we do for these ripple tank photographs. Instead, we observe only where the waves strike some object, and from this deduce the nature of the waves.

To illustrate what we mean, imagine a harbor with a sea wall and two narrow entrances in the wall as shown in Figure (10). Waves coming in from the ocean emerge as circular waves from each entrance and produce a two slit interference pattern in the harbor. Opposite the sea wall is a beach as shown.

If we are at point A on the beach directly across from the center of the two entrances, we are standing in the center beam of waves in the interference pattern. Here



#### Figure 9

The wave pattern emerging from 2 slits is similar to the wave pattern produced by two plungers.

large waves wash up on the beach. Walking north along the beach we cross the first line of nodes at point B. Here the water is calm. Going farther up to point C we are again in the center of a beam of waves. We will call this the *first maximum* above the *central maximum*. Farther up we cross the second line of nodes at point D and encounter the second minimum in the height of waves striking the shore.

Going south from point A we encounter the same alternate series of maxima and minima at points B', C', D', etc. If we graphed the amplitude of the waves striking the shore, we would get the pattern shown at the right side of Figure (10).

Now suppose that we walk along the beach on a calm day where there are no waves, but on the previous day there had been a storm. During the storm, the waves striking the shore eroded the beach. As you walk along the beach you notice a series of indentations, at points A, C, C', etc. where the beach was eroded. The sand was not eroded at points B, B', D and D'. If someone asked what the ocean waves were like during the storm, could you tell them?

By measuring the distance between the maximum erosions and knowing the geometry of the harbor, you can determine the wavelength of the ocean waves that struck the sea wall during the storm. Similar calculations can be made to determine the wavelength of any kind of wave striking two narrow slits producing an interference pattern on the other side. We do not have to see the actual wave pattern, we only have to note the location of the maxima and minima of the waves striking an object like the shore in Figure (10).



#### Figure 10

Hypothetical harbor with two entrances through the sea wall. If ocean waves are coming straight in toward the sea wall, there will be a 2 slit interference pattern inside the harbor, with a series of maxima and minima along the beach.

We begin our analysis of the two slit wave pattern by drawing a series of circles to represent the wave crests and troughs emerging from the two slits. The results, which are shown in Figure (11), are essentially the same as our analysis of the two plunger interference pattern in Figure (2). The maxima occur where crests meet crests and troughs meet troughs. The minima or lines of nodes are where crests meet troughs.

# Exercise 2

On Figure (11), sketch the lines along which crests meet troughs, i.e., where solid and dashed circles intersect. This should be where the lines of nodes are located.

# **The First Maxima**

The central maximum is straight across from the center of the two slits (if the incoming waves are parallel to the slits as in Figure 11). To figure out where the first maximum is located, consider the sketch in Figure (12). We have reduced the complexity of the sketch by drawing only the solid circles representing wave crests. In addition we have numbered the crests emerging



## Figure 11

Analysis of the two slit wave pattern, assuming that circular waves emerge from each slit and interfere with each other. The maxima are where crests from one slit meet crests from the other. Cancellation occurs where crests meet troughs.





from each slit. We see that at first maximum, the 12th crest from the lower slit has run into the 11th crest from the upper slit, producing a maximum crest. *The distance from the lower slit to the first maximum is exactly one wavelength longer than the distance from the upper slit to the first maximum.* This is what determines the location of the first maximum.

In Figure (13) we have repeated the sketch of Figure (12), but now focus our attention on the difference in the length of the two paths from the slits to the first maximum. Since an extra wavelength  $\lambda$  fits into the lower path, the path length difference is  $\lambda$  as shown. The bottom path, with  $\lambda$  removed, and the upper path, both shown as dashed lines in Figure (13), are thus the same length and therefore form 2 sides of an isosceles triangle.

Let us denote by  $\theta_1$  the angle from the center of the two slits up to the first maximum. Since this line bisects the isosceles triangle formed by the two dashed lines, it is perpendicular to the base of the isosceles triangle which is the line from the center of the upper slit down to the point (a) on the lower path. As a result, the base of the isosceles triangle makes the same angle  $\theta_1$  with the plane of the slits as the line to the first maximum does with the horizontal line to the central maximum. (Picture rotating

the isosceles triangle up around its base. If you rotate the isosceles triangle by an angle  $\theta_1$ , its base will rotate by the same angle  $\theta_1$ , thus the 2 angles labeled  $\theta_1$  in Figure (13) are the same.)

Our approximation in this analysis is that the separation d between the slits is very small compared to the distance over to where we are viewing the first maximum. If this is true, then the two paths to the first maximum are essentially parallel and the small bold triangle in Figure (13) is very nearly a right triangle. Assuming that this is a right triangle, we immediately get

$$\sin \theta_1 = \frac{\lambda}{d} \qquad \begin{array}{c} angle \ to \ first \\ maximum \end{array} \tag{1}$$

In Figure (14) we have another right triangle involving the angle  $\theta_1$ . If the distance from the slits to where we are viewing the maxima is D, and if we designate by  $Y_{max}$  the distance from the central to the first maximum, then the hypotenuse of this right triangle is given by the Pythagorean theorem as  $\sqrt{D^2 + Y_{max}^2}$ . From this triangle we have

$$\sin \theta_1 = \frac{Y_{max}}{\sqrt{D^2 + Y_{max}^2}}$$
(2)

Equating the two formulas for sin  $\theta_1$  and solving for  $\lambda$  gives

$$\lambda = Y_{max} \frac{d}{\sqrt{D^2 + Y_{max}^2}}$$
(3)



**Figure 13** The path length difference to the first maximum is one wavelength  $\lambda$ .

An easy way to remember this derivation is to note that the two triangles in Figures (13) and (14), drawn separately in Figure (15), are similar triangles. Thus the ratios of the small sides to the hypotenuses must be equal, giving

$$\frac{\lambda}{d} = \frac{Y_{\text{max}}}{\sqrt{D^2 + Y_{\text{max}}^2}}$$
(3')

The importance of Equation 3 is that it allows us to calculate the wavelength of a wave by observing the distance  $Y_{max}$  between maxima of the interference pattern. For example, in our problem of determining the character of the waves eroding the beach in Figure (10), we could use a map to determine the distance D from the breakwater to the shore and the distance d between entrances through the breakwater. Then pacing off the distance  $Y_{max}$  between erosions on the beach, we could use Equation 3 to determine what the wavelength of the waves were during the storm.



#### Figure 14

The angle  $\theta_1$  up to the first maxima is the same as the angle in the small triangle of Figure (13).



Figure 15 Since the two triangles are similar, we have  $\lambda/d = y_{max} / \sqrt{D^2 + Y_{max}^2}$ 

# **Exercise 3**

Repeat the derivation that led to Equation 3 except do the calculation in terms of the distance  $Y_{min}\,$  from the central maximum to the first minimum. (Now the path length difference is  $\lambda/2$ .)

# **Exercise 4**

We will see that Equation 3 has an applicability that goes far beyond the analysis of two slit interference patterns. You will need this formula several times later in this course, and quite likely in other research work. Rather than memorizing the formula, it is much better to memorize the derivation. The best way to do this is to treat the derivation as a clean desk problem. Some time, a day or so after you have read this section, clean off your desk, take out a blank sheet of paper, and derive Equation 3. The first time you try it, you may have forgotten some steps. If that happens, review the derivation and try to do the clean desk problem a day or so later. It is worth the effort because the derivation summarizes all the formulas used in this chapter.

# **TWO SLIT PATTERN FOR LIGHT**

Christian Huygens discovered his principle of wave motion in 1678, and developed a wave theory of light that competed with Newton's particle theory of light. It was not until 1801, over 120 years later, that Thomas Young first demonstrated the wave nature of light using a two slit interference experiment. Why did it take so long to do this demonstration?

Two major problems arise when you try to test for the wave nature of light. One is the fact that the wavelength of light is very short, on the order of one hundred thousand times shorter than the wavelengths of the water waves we observe in the ripple tank photographs. A more serious problem is that individual atoms in the sun or a light bulb emit short bursts of light that are not coordinated with each other. The result is a chopped up, incoherent beam of light that may also include a mixture of frequencies.

In our analogy of a sea wall with two entrances, it is likely that a real storm would produce a mixture of waves of different wavelengths heading in different directions. Many different interference patterns would be superimposed on the inside of the sea wall, different maxima and minima would overlap at the beach and the beach would be more or less uniformly eroded. Walking along the beach the next day, you would not find enough evidence to prove that the damage was done by ocean waves, let alone trying to determine the wavelength of the waves.

The invention of the laser by Charles Townes in 1960 eliminated the experimental problems. The laser emits a continuous coherent beam of light that more closely resembles the orderly ripple tank waves approaching the slits in Figure (10) than the confused wave motion seen in a storm. If you send a laser beam through two closely spaced slits, you cannot help but see a two slit interference pattern. Even in a demonstration lecture, the two slit pattern produced by a laser beam can be used to measure the wavelength of the light in the beam. In Figure (16) we placed a two slit mask next to a millimeter scale on the top of an overhead projector and projected the image on a large screen. You can see that the spacing between the two slits is about 1/3 of a millimeter. In Figure (17) we aimed the red beam of a common helium neon laser through the two slits of Figure (16), onto a screen 10 meters from the slits. The resulting two slit pattern consisting of the alternate maxima and minima are easily seen by the class. Marking the separation of two maxima on a piece of paper and measuring the distance we found that the separation  $Y_{max}$  between maxima was about 2.3 cm.

In using Equation 3,  $\lambda = Y_{max}d/\sqrt{D^2 + Y_{max}^2}$  to calculate the wavelength  $\lambda$ , we note that the 10 meter distance D is much greater than the 2.3 cm  $Y_{max}$ . Thus we can

neglect the  $Y^2_{max}\,$  in the square root and we get the simpler formula

$$\lambda \approx Y_{\max} \frac{d}{D} \quad (if D >> Y_{\max})$$
 (3a)

Putting in the numbers obtained from Figures (16) and (17), we get

$$\lambda = 2.3 \text{ cm} \times \frac{.3 \times 10^{-3} \text{m}}{10 \text{m}} = 7 \times 10^{-5} \text{cm}$$
 (4)

While this demonstration experiment gives fairly approximate results, accurate to about one significant figure, it may be somewhat surprising that a piece of apparatus as crude as the two slits seen in Figure (16) even allows us to measure something as small as  $7 \times 10^{-5}$  cm.



#### Figure 16

The two slits and a plastic ruler are placed on an overhead projector and projected onto a screen 10 meters away. This is a photograph of the screen.



#### Figure 17

The 2 slit laser pattern is then projected on the screen. Below is a centimeter scale, showing that the maxima are about 2.3 centimeters apart.

Shows that d = .3mm

Shows that  $Y_{max} = 2.3 \text{ mm for } D = 10 \text{ m}.$ 

# THE DIFFRACTION GRATING

The crudeness of our measurement of the wavelength of the laser light in our two slit experiment could be improved somewhat by a more accurate measurement of the separation of the two slits, but the improvement would not be great. There is, however, a simple way to make far more accurate measurements of the wavelength of a beam of light. The trick is simply add more slits.

To see why adding more slits gives more accurate results, we show in Figure (18) the wave patterns we get when the laser beam is sent through two slits, three slits, four slits, five slits, and seven slits. We created the slits using a Macintosh computer using the Adobe Photoshop program and a Linatronic printer to produce the film images of the slits. The Linatronic printer can draw precise lines one micron wide  $(10^{-6} \text{ meters})$ ; thus we

had excellent control over the slit width and spacing. For these images, the slits are 50 microns ( $50\mu$ ) wide and spaced 150 microns apart on centers.

The photographs of the interference patterns produced by the slits of Figure (18) are all enlarged to the same scale. The important point to notice is that while the maxima become sharper as we increase the number of slits, the spacing between maxima remains the same. *Adding more identical slits sharpens the maxima but does not change their spacing!* As a result the two slit formula, Equation 3, can be applied to any number of slits as long as the spacing d between slits remains constant.

If there are many slits, the device is called a *diffraction grating* and Equation 3, which we repeat below, is known as the diffraction grating formula.



## Figure 18

Interference patterns for various slit structures. If we keep the spacing between slits the same, then there is no change in the location of the maxima, no matter how many slits the laser beam passes through. Thus an analysis of the location of the maxima for 2 slits applies to any number of slits. Also note that the single slit pattern acts as an envelope for the multiple slit patterns.



#### **Exercise 5**

In Figure (18) the separation of the slits is 150 microns and the separation of 10 maxima is 26.4 cm. The screen is a distance of 6.00 meters from the slits. From this determine the wavelength of the light in the laser beam

- (a) using the exact formula, Equation 3.
- (b) using the approximate formula, Equation 3a.

How many significant figures are meaningful in your result? To this accuracy, did it make any difference whether you used the exact Equation 3 or the approximate Equation 3a.

Figure (18) demonstrates that the more slits you use, the sharper the maxima and the more accurately you can determine the wavelength of the light passing through the slits. In the latter part of the 1800s, the diffraction grating was recognized as an excellent tool for scientific research, and a great effort was put into producing gratings with as many closely spaced lines as possible. Fine ruling machines were developed that produced gratings on the order of 6000 lines or slits per centimeter. With so many lines, very sharp maxima are



## **Exercise 6**

In Figure (19), a laser beam is sent through a smoke filled box with a diffraction grating at the center of the box as shown in the sketch (19a). The smoke allows you to see and photograph the central laser beam and two maxima on each side. You also see maxima reflected from the back side of the grating. (When you shine a laser beam on a CD you get only the reflected maxima, no light goes through the record.)

The grating used in Figure (19) had 15,000 lines per inch (1 inch = 2.54 cm). From this information and the photograph of Figure (19b), determine the wavelength of the laser beam used. Try both the exact Equation 3 and the approximate Equation 3a. Explain why Equation 3a does not work well for this case.



#### Figure 19

Laser beam passing through a diffraction grating. The beam is made visible by placing the grating in a smoke filled box. Because the lines are so close together, the maxima are widely separated. You can also see reflected maxima on the back side.



# More About Diffraction Gratings

The results of Figure (18) demonstrated that the maxima got sharper but remained in the same place as we added slits. Let us now see why this happens.

The maxima of a diffraction grating occur at those points on the screen where the waves from every slit add up constructively. This can happen only when the path length difference between neighboring maxima is 0 (central maximum),  $\lambda$  (first maxima),  $2\lambda$  (second maxima), etc. In Figure (20) we are looking at a small section of a diffraction grating where we have drawn in the paths to the first maxima. The path length differences between neighboring slits are all  $\lambda$  and the angle  $\theta_1$  to the first maxima is given by  $\theta_1 = \lambda/d$ , the same results we had for the two slit problem in Figure (13). This angle does not depend upon the number of slits, thus the position of the maxima do not change when we add slits as in Figure (18).

To see why the maxima become narrower as we add slits, let us consider the example of a 1000 slit grating illustrated in Figure (21). We have numbered the slits from 1 to 1000, and are showing the paths to a point just below the first maximum where the path length difference between neighboring slits is  $(\lambda - \lambda/1000)$  instead of  $\lambda$ .

On the figure we are indicating, not the path length difference between neighboring slits, but instead, the path length difference between the first slit and the others. This difference is  $(\lambda - \lambda/1000)$  for slit #2,  $(2\lambda - 2\lambda/1000)$  for slit #3,  $(3\lambda - 3\lambda/1000)$  for slit #4, etc.



#### Figure 20

When the path length difference between neighboring paths is  $\lambda$ , then the waves from all slits add constructively and we get the first maxima.

When we get down to slit # 501, just over half way length difference down, the path is  $500\lambda - 500\lambda/1000 = 500\lambda - \lambda/2$ . In other words, the waves from slit 1 and slit 501 are precisely one half a wavelength out of phase, crests exactly meet troughs, and there is precise cancellation. A similar argument shows that waves from slits #2 and #502 are  $\lambda/2$  out of phase and cancel exactly. The same goes for the pairs #3 and #503, #4 and #504, all the way down to 500 and 1000. In other words, the waves all cancel in pairs and we have a minimum, complete cancellation at the point just below the first maximum where the path length difference is  $\lambda - \lambda/1000$  instead of  $\lambda$ . With two slits we got complete cancellation half way between maxima. With 1000 slits, we only have to go approximately 1/1000 the way toward the next maxima before we get complete cancellation. The maxima are roughly 500 times sharper. You can see that with n slits, the maxima will be about n/2 times sharper than for the two slit example.



#### Figure 21

In a thousand slit grating, we get complete cancellation when the path length difference between neighboring slits is reduced from  $\lambda$  to  $\lambda - \lambda/1000$ .

The maxima will also be much more intense because the light is coming in from more slits. If we have n slits, the amplitude of the wave at the center of a maxima will be n times as great as the amplitude from a single slit. It turns out that the amount of energy in a wave, the intensity, or, for light, the brightness, is proportional to the square of the amplitude of the wave. Thus the brightness at the center of the maxima for an n slit grating is  $n^2$  times as bright at the brightness we would have for a single slit. The maxima for the 1000 slit grating illustrated in Figure (21) would be one million times brighter than if we let light go through only one of the slits. (To see how the total energy works out, consider the following argument. Compared to one slit, when you have n slits, you have n times as much light energy that is compressed into a maxima that is only 1/n as wide. You get one factor of n in brightness due to the compression, and the other factor of n due to there being n slits.)



#### Figure 22

When white light passes through a diffraction grating, the maxima for different colors emerge at different angles. Since red light has the longest wavelength of the visible colors, it emerges at the greatest angle.

#### The Visible Spectrum

Thus far we have been using a laser beam to study the operation of a diffraction grating. Now we will reverse the process and use diffraction gratings to study the nature of beams of light.

If you send a beam of white light through a diffraction grating, you get a series of maxima. In all but the central maxima light is spread out into a rainbow of colors illustrated in Figure (22). In each maxima the red light is bent the most, and blue the least. As we saw from Equation 1,  $\sin \theta_1 = \lambda/d$ , the longer the wavelength the greater the angle the wave is bent or diffracted. Thus red light has the longest wavelength and blue the shortest in the mixture of wavelengths that make up white light.

The longest wavelength that the human eye can see is about  $7.0 \times 10^{-5}$  cm, a deep red light, and the shortest is about  $4.0 \times 10^{-5}$  cm, a deep purple. All other visible wavelengths, the entire spectrum of visible light, lies in the range between  $4.0 \times 10^{-5}$  cm to  $7.0 \times 10^{-5}$  cm. Yellow light, for example, has a wavelength around  $5.7 \times 10^{-5}$  cm, and green light is near  $5.0 \times 10^{-5}$  cm.

As we saw in Chapter 32, visible light is just a part of the complete electromagnetic spectrum. A surprisingly small part. As radio, television, microwave ovens, infra red sensors, ultraviolet sunscreens, x ray photographs, and  $\gamma$  ray bursts in the sky have entered our experience of the world, we have become familiar with a much greater range of the electromagnetic spectrum. As indicated in Figure (23), AM radio waves have wavelengths in the range of 10 to 100 meters, VHF television a few meters, VHF from around 10 cm to a meter, microwaves from around a millimeter to 10 cm, infra red from less than a millimeter down to visible red light at  $7.0 \times 10^{-5}$  cm. At shorter wavelengths



Visible light is a tiny piece of the electromagnetic spectrum.

than deep blue we have ultraviolet, then x rays, and the very shortest wavelengths are called  $\gamma$  (gamma) rays.

To study the electromagnetic spectrum, different devices are used at different wavelengths. In Chapter 32 we used a loop of wire and Faraday's law to detect the magnetic fields of a radio wave. This required the use of an oscilloscope that could display radio wave frequencies, typically of the order of a megacycle for AM radio. For visible light the frequencies are too high, the wavelengths too short for light to be studied by similar techniques. Instead the diffraction grating will be our main tool for studying the electromagnetic waves in the visible spectrum.

# **Exercise 7**

What are the lowest and highest frequencies of the waves in the visible spectrum? What is the color of the lowest frequency? What is the color of the highest? What is the frequency of yellow light?

# **Atomic Spectra**

Our main application of the diffraction grating will be to study the spectrum of light emitted by atoms. It has long been known that if you have a gas of a particular kind of atom, like nitrogen, oxygen, helium, or hydrogen, a special kind of light is emitted. You do not get the continuous blend of wavelengths seen in white light. Instead the light consists of a mixture of distinct wavelengths. Which wavelengths are involved depends upon the kind of atom emitting the light. The mixture of wavelengths provide a unique signature of that atom, better than a fingerprint, for identifying the presence of an atom in a gas. In fact, the element helium (named after the Greek word helios for sun) was first identified in the sun by a study of the spectrum of light from the sun. Only later was helium found here on earth.

The subject of modern astronomy is based on the study of the spectrum of light emitted by stars. Some stars consist mostly of hydrogen gas, others a mixture of hydrogen and helium, while still others contain various amounts of heavier elements. We learn the composition of the star by studying the spectrum of light emitted, and from the composition we can deduce something about the age of the star and the environment in which it was formed.

Our main reason for studying the spectrum of light emitted by atoms will be to learn something about the atoms themselves. Since Rutherford's discovery of the atomic nucleus in 1912, it has been known that atoms consist of a positively charged nucleus surrounded by negatively charged electrons. If we apply Newtonian mechanics to predict the motion of the electrons, and Maxwell's equations to predict the kind of electromagnetic radiation the moving electrons should radiate, we get the wrong answer. There is no way that we can explain the spectrum of light emitted by atoms from Maxwell's equations and Newtonian mechanics. The existence of detailed atomic spectra is a clue that something is wrong with this classical picture of the atom. It is also the evidence upon which to test new theories.

We do not have to study many kinds of atoms to find something wrong with the predictions of classical theory. The simplest of all atoms, the hydrogen atom consisting of one proton for a nucleus, surrounded by one electron, is all we need. Heated hydrogen gas emits a distinct, orderly, spectrum of light that provides the essential clues of what is going on inside a hydrogen atom. In this chapter we will focus on using a diffraction grating to learn what the spectrum of hydrogen is. In the following chapters we use the hydrogen spectrum to study the atom itself.



**Figure 24** *Apparatus to measure hydrogen spectrum.* 

# THE HYDROGEN SPECTRUM

The apparatus required for studying the hydrogen spectrum can be as simple as the hydrogen source, meter stick and diffraction grating shown in the photograph of Figure (24). The hydrogen source consists of a narrow glass tube filled with hydrogen gas, with metal electrodes at the ends of the tube. When a high voltage is applied to the electrodes, an electric current flows through the gas, heating it and causing it to emit light. The diffraction grating is placed in front of the hydrogen tube, and the meter stick is used to measure the location of the maxima.

The setup of the apparatus is illustrated in Figure (25) and the resulting spectrum in Figure (26). In this spectrum we are looking at the first maxima on the left side of the meter stick as shown in Figure (25). The leftmost line, the one bent the farthest is a deep red line which is called the *hydrogen*  $\alpha$  line, and labeled by  $\alpha$  in the photograph. The next line is a spurious line caused by impurities in the hydrogen tube. More to the right is a bright, swimming-pool blue line called *hydrogen*  $\beta$ . Much harder to see is the third line called *hydrogen*  $\gamma$ , a deep violet line near the short wavelength end of the visible spectrum. The three lines  $\alpha$ ,



#### Figure 25

To determine the wavelength of light using a diffraction grating, you need to measure the distance  $Y_{max}$  to the first maximum, and the distance D shown. To measure  $Y_{max}$ , slide the pointer along the meter stick until it lines up with the first maximum.  $\beta$  and  $\gamma$  are the only lines emitted by pure hydrogen gas in the visible part of the electromagnetic spectrum. Their wavelengths are

$$\lambda_{\alpha} = 6.56 \times 10^{-5} \text{cm}$$
  

$$\lambda_{\beta} = 4.86 \times 10^{-5} \text{cm}$$
  

$$\lambda_{\gamma} = 4.34 \times 10^{-5} \text{cm}$$
(5)

When actually performing the experiment shown in Figure (24), there are some steps one should take to improve the accuracy of the results. As shown in Figure (27), a small arrowhead is placed on the grating itself. You then place your eye behind the meter stick and move your head and the slider on the meter stick until the point on the slider lines up with the arrowhead on the grating and with the spectral line you are trying to measure.



#### **Figure 26** Photograph of the $\alpha$ , $\beta$ and $\gamma$ lines in the hydrogen spectrum.



**Figure 27** Looking through the grating, move your eye so that the spectral line is centered over the pointer as shown.

Rather than trying to measure the distance  $Y_{max}$  from the central maximum to the spectral line, it is more accurate to measure the distance  $2Y_{max}$  from the first maximum on the left to the first maximum on the right, and then divide by 2. The wavelength of the line under study is then given by the diffraction grating formula, Equation 3

$$\lambda = Y_{max} \frac{d}{\sqrt{D^2 + Y_{max}^2}}$$
 (3 repeated)

where d is the separation of the slits in the grating and D the distance from the grating to the meter stick.

(When you first perform this experiment you may be confused by where the central maximum is. If you look straight through the grating at the tube, all you see is the tube. But that is the central maximum. It looks like the tube because all the colors go straight through the grating. There is no separation of colors or distortion of the image. To see a spectrum you have to look through the grating but far off to the side from the tube.)

## **Exercise 8**

Derive a formula for the wavelength  $\lambda$  of a spectral line in terms of the distance  $Y_{2\,max}$  from the central maximum to the second maxima of the line. The second maxima of the bright lines of an atomic spectra are quite easily seen using the apparatus of Figure (24).

# The Experiment on Hydrogen Spectra

You should carry out the following steps when doing the hydrogen spectrum experiment shown in Figure (25).

(1) Determine the wavelength of all the spectral lines you can see, and compare your results with those given in Equation 5. Measure distances between first maxima, not to the central maxima.

(2) Measure the distances to the second maxima for the lines you can see out there and compute the corresponding wavelengths using our results from Exercise 8. Compare these wavelengths with those you get using the first maxima.

## **The Balmer Series**

There are many spectral lines emitted by the hydrogen atom. Only three, however, are in the visible part of the spectrum. The complete spectrum consists of a number of series of lines, and the three visible lines belong to the series called the *Balmer series*. The red line, hydrogen  $\alpha$ , is the longest wavelength line in the Balmer series, next comes the blue hydrogen  $\beta$ , then the violet hydrogen  $\gamma$ . Then there are many lines of the Balmer series out in the ultraviolet, which we cannot see by eye, but which we can record on photographic film.

Figure (28) shows part of the spectrum of light from a hydrogen star. These lines are in the ultraviolet and are all part of the Balmer series. Slightly different naming is used here. In the notation of Figure (28), we should call the red hydrogen  $\alpha$  line H3, the blue  $\beta$  line H4, and the violet  $\gamma$  line H5. In Figure (28), the first 6 Balmer lines are missing. Here we see lines H9 through H40. As the lines increase in number they get closer and closer together. The whole series ends with very many, very closely spaced lines near  $3.65 \times 10^{-5}$  cm. It is called a *series* because the lines converge to a final wavelength in much the same way that many mathematical series converge to a final value.

It was the Swiss school teacher Johann Balmer who in 1885 discovered a formula for the wavelengths of the spectral lines seen in Figure (28). The wavelength of the m th line (m=3 for H3, m=4 for H4, etc.) is given by the formula

$$\lambda_{\rm m} = 3.6456 \times 10^{-5} \,{\rm cm} \times \left(\frac{{\rm m}^2}{{\rm m}^2 - 4}\right)$$
 (6)

Equation 6 is known as the *Balmer formula*.

For m=3 we get from the Balmer formula

$$\lambda(H3) = 3.6456 \times 10^{-5} \text{ cm} \times \left(\frac{9}{9-4}\right)$$
  
= 6.56 × 10^{-5} \text{ cm} (6a)

which agrees with Equation 5 for hydrogen  $\alpha$ . Each higher value of m gives us the wavelength of a new line. At large values of m the factor  $m^2/(m^2 - 4)$  approaches 1, and the lines get closer and closer together as seen in Figure (28). The end is at  $3.65 \times 10^{-5}$  cm where m is very large.

## **Exercise 9**

(a) Use Equation 6 to calculate the wavelengths of the  $\beta$  and  $\gamma$  lines of the hydrogen spectrum and compare the results with Equation 5.

(b) Calculate the wavelength of H40 and compare your results with Figure (28).



#### Figure 28

Spectrum of the star HD193182, showing ultra violet hydrogen lines near the limit of the Balmer series. This series of lines begins in the visible part of the spectrum with the lines we have called  $\alpha$ ,  $\beta$ , and  $\gamma$ , (which would be called H3, H4, and H5 in this diagram), and goes on to the ultra violet. The lines get closer and closer together, until the end just beyond the point labeled H40. The Swiss school teacher Johann Balmer discovered a formula for the wavelengths of these lines.

# THE DOPPLER EFFECT

One phenomena of wave motion that is particularly easy to visualize is the *Doppler effect*. As you can see in Figure (29), if the wave source is moving, the wavelength of the waves is compressed in front of the source and stretched out behind. This result, which is obvious for water waves, also applies to sound waves in air and to light waves moving through space.

To analyze the effect, we first note that if the source is at rest, then the waves all travel out from the source at a speed  $v_{wave}$ , have a wavelength  $\lambda_0$  and a period  $T_0$  given by

$$T_0 \frac{\sec}{\text{cycle}} = \frac{\lambda_0 \text{ cm/cycle}}{v_{\text{wave}} \text{ cm/sec}} = \frac{\lambda_0}{v_{\text{wave}}} \frac{\sec}{\text{cycle}}$$
(7)

If the source is moving forward at a speed  $v_{source}$ , then during one period  $T_0$  the source will move forward a distance  $x = v_{source}T_0$ . But this is just the amount  $\Delta\lambda$ by which the wavelength is shortened in front and stretched out in back. Thus

$$\Delta \lambda = v_{\text{source}} T_0 = v_{\text{source}} \frac{\lambda_0}{v_{\text{wave}}}$$
(8)

where we used Equation 7 to replace  $T_0$  by  $\lambda_0/v_{wave}$ .



#### Figure 29

When the source of the wave is moving, the wavelengths are compressed in front and stretched out behind.

As a result the wavelengths in front and back of the source are

$$\lambda_{\text{front}} = \lambda_0 - \Delta \lambda = \lambda_0 \left( 1 - \frac{v_{\text{source}}}{v_{\text{wave}}} \right)$$
 (9a)

$$\lambda_{\text{back}} = \lambda_0 + \Delta \lambda = \lambda_0 \left( 1 + \frac{v_{\text{source}}}{v_{\text{wave}}} \right)$$
 (9b)

If we are in front of the moving source, the wave period  $T_{front}$  we observe is the time it takes the shortened wavelength  $\lambda_{front}$  to pass us at a speed  $v_{wave}$ , which is

$$T_{\text{front}} = \frac{\lambda_{\text{front}}}{v_{\text{wave}}} = \frac{\lambda_{0}}{v_{\text{wave}}} \left(1 - \frac{v_{\text{source}}}{v_{\text{wave}}}\right)$$
$$T_{\text{front}} = T_{0} \left(1 - \frac{v_{\text{source}}}{v_{\text{wave}}}\right)$$
(10a)

where we now replaced  $\lambda_0/v_{wave}$  by T<sub>0</sub>. In the back, the period is extended to

$$T_{back} = T_0 \left( 1 + \frac{v_{source}}{v_{wave}} \right)$$
(10b)



#### Figure 30

When the source is moving faster than the waves, the waves build up on the front edge to create a shock wave. For supersonic flight, this shock wave produces the sonic boom. If the speed of the source approaches the speed of the wave, as in the case of a jet airplane approaching the speed of sound, the wavelength in front goes to zero. At speeds greater than the speed of the wave, as in supersonic flight, there are no waves ahead of the source; instead, the leading edge of the waves pile up as shown in Figure (30) to create what is called a *shock wave*. This shock wave is responsible for the sonic boom we hear when a jet passes overhead at supersonic speeds.

## Exercise 10

There is a simple experiment you can perform to observe the Doppler effect. Stand beside a road and have a friend drive by at about 40 mi/hr while blowing the car horn. As the car passes, the pitch of the horn will suddenly drop because the wavelength of the sound waves, which was shortened as the car approached, is lengthened after it passes. The shorter, higher-pitched sound waves change to longer, lower-pitched waves.

For this exercise, assume the car is owned by a musician, and the car horn plays the musical note A at a frequency of 440 cycles per second.

*a)* What is the wavelength of a 440 cycle/sec note, if the speed of sound is 1000 ft/sec?

*b)* What is the wavelength of the note we hear if the car is approaching at a speed of 40 miles/hr?

*c)* What is the frequency we hear if the car is approaching at 40 miles/hr?

*d)* What is the frequency we hear when the car is going away from us at 40 miles/hr?

# Stationary Source and Moving Observer

If the source is at rest but we, the observer, are moving, there is also a Doppler effect. In the case of water or sound waves, if we are moving through the medium toward the source, then the wave crests pass by us at an increased relative speed  $v_{rel} = v_{wave} + v_{us}$ . Even though the wavelength is unchanged, the increased speed of the wave will carry the crests by faster, giving us an apparently shorter period and higher frequency.

If our velocity through the medium is small compared to the wave speed, then we observe essentially the same decrease in period and increase in frequency as in the case when the source was moving. In particular, Equation 10 is approximately correct.

On the other hand, when the waves are in water or air and the relative speed of the source and observer approaches or exceeds the wave speed, there can be a considerable difference between a moving source and a moving observer. As illustrated in Figure (31a), if the source is moving faster than the wave speed, there is a shock wave and the observer detects no waves until the source passes. But if the source is at rest as in Figure (31b), there is no shock front and the observer moves through waves before getting to the source, even if the observer is moving faster than the wave speed.



## Figure 31

For waves in water or air, there can be a significant difference between a moving source with a stationary observer, and a moving observer with a stationary source, even though the source and observer have the same relative velocity in the two cases. For light, the principle of relativity requires that the two cases be identical.

# **Doppler Effect for Light**

When a source of light waves is moving toward or away from us, there is also a Doppler effect. If the source is moving toward us, the wavelengths we see are shortened. This means that the color of the light is *shifted toward the blue*. If the source is moving away, the wavelengths are stretched out, become longer, and the color *shifts toward the red*. When the speed of the source is considerably less than the speed of light, Equations 9 and 10 correctly give the observed wavelength  $\lambda$  and period T in terms of the source's wavelength  $\lambda_0$  and period and T<sub>0</sub>.

#### **Principle of Relativity**

There is one fundamental difference, however, between the Doppler effect for water and sound waves, and the Doppler effect for light waves. For water and sound waves we could distinguish between a source at rest with a moving observer and an observer at rest with a moving source. If the source were at rest, it was at rest *relative to the medium through which the wave moves*. We got different results depending on whether it was the source or the observer that was at rest.

In the case of light, the medium through which light moves is *space*. According to the principle of relativity, one cannot detect uniform motion relative to space. Since it is not possible to determine which one is at rest and which one is moving, we must have exactly the same Doppler effect formula for the case of a stationary source and a moving observer, or vice versa. The Doppler effect formula can depend only on the *relative velocity* of the source and observer.

One way to use the principle of relativity is to always assume that you yourself are at rest relative to space. (No one can prove you are wrong.) This suggests that we should start from Equations 9 and 10, which were derived for a stationary observer, and replace  $v_{wave}$  by the speed of light c, and interpret  $v_{source}$  as the relative velocity between the source and the observer. Equations 9 and 10, modified this way, are correct as long as the source is not moving too fast. However if the source is moving relative to us at a speed approaching the speed of light, there is one more relativistic effect that we have to take into account. Remember that a moving clock runs slow by a factor of  $\sqrt{1 - v^2/c^2}$ . If the source is radiating a light wave of period  $T_0$ , then that period can be used in the construction of a clock. If we observe the source go by at a speed  $v_{source}$ , the period  $T_0$  must appear to us to increase to  $T_0'$  given by

$$T_0' = \frac{T_0}{\sqrt{1 - v_{source}^2/c^2}}$$
 (see Eq 1-11)

From our point of view, the source is radiating light of period  $T_0'$ . This is the light whose wavelength is stretched or compressed, depending on whether the source is moving away from or towards us. Thus we should use  $T_0'$  instead of  $T_0$  in Equation 10.

Replacing  $T_0$  by  $T_0'$  in Equation 10 gives

$$T_{\text{front}} = \frac{T_0}{\sqrt{1 - v_{\text{source}}^2/c^2}} \left(1 - \frac{v_{\text{source}}}{c}\right) \quad (11a)$$
$$T_{\text{back}} = \frac{T_0}{\sqrt{1 - v_{\text{source}}^2/c^2}} \left(1 + \frac{v_{\text{source}}}{c}\right) \quad (11b)$$

where  $v_{source}$  is the speed of the source relative to us, and we have set  $v_{wave} = c$ . Equations (11) are the relativistic Doppler effect equations for light. They are applicable for any source speed, even if the source is moving relative to us at speeds approaching the speed of light. The corresponding wavelengths are

$$\lambda_{\text{front}} = c \frac{cm}{sec} \times T_{\text{front}} \frac{sec}{cycle} = cT_{\text{front}} \frac{cm}{cycle}$$

$$\lambda_{\text{front}} = cT_{\text{front}}$$

$$\lambda_{\text{back}} = cT_{\text{back}}$$
(12)

## Exercise 11

Using Equations (12), express  $\lambda_{front}$  and  $\lambda_{back}$  in terms of  $\lambda_0,$  v\_{source} and c.

# Exercise 12

In Figure 31b, where we picture a stationary source and a moving observer, the waves pass by the observer at a speed  $v_{wave} + v_{observer}$ . Why can't this picture be applied to light, simply replacing  $v_{wave}$  by c and letting  $v_{observer}$  be the relative velocity between the source and observer.

# **Doppler Effect in Astronomy**

The Doppler effect has become one of the most powerful tools astronomers use in the study of the universe. Assuming that distant stars and galaxies are made up of the same matter as nearby stars, we can compare the spectral lines emitted by distant galaxies with the corresponding spectral lines radiated by elements here on earth. A general shift in the wavelengths to the blue or the red, indicates that the source of the waves is moving either toward or away from us. Using Equations 12 we can then quite accurately determine how fast this motion toward or away from us is.



#### Figure 32

The most distant galaxy observed as of January 1995. This galaxy, given the romantic name 8C 1435+63, was photographed by the Keck telescope in Hawaii. The two halves of the distant galaxy are indicated by the white bracket. The galaxy is moving away from us at 95% the speed of light. Until the 1960s astronomers did not have much need for the relativistic Doppler shift equations. The non relativistic Equations 10 were generally adequate because we did not observe stars or galaxies moving relative to us at speeds greater than 10 to 20% the speed of light. But that changed dramatically with the discovery of quasars in 1963. Quasars are now thought to be brilliant galaxies in the early stages of formation. They can be seen from great distances and are observed to move away from us at speeds as great as 95% the speed of light. To analyze such motion, the relativistic formulas, Equations 11 and 12 are clearly needed.

# Exercise 13

The most rapidly receding galaxy observed by the spring of 1995 is the galaxy named 8C 1435 + 63 shown in the photograph of Figure (32) taken by the Keck telescope in Hawaii. Much of the light from this galaxy is radiated by hydrogen gas. This galaxy is moving away from us at a speed  $v_{source} = .95c$ .

(a) Assuming that the hydrogen in this galaxy radiates the same spectrum of light as the hydrogen gas in our discharge tube of Figure (24), what are the wavelengths of the first three Balmer series lines  $\lambda_{\alpha}$ ,  $\lambda_{\beta}$ , and  $\lambda_{\gamma}$ , by the time these waves reach us. (They will be greatly stretched out by the motion of the galaxy.)

(b) Astronomers use the letter z to denote the relative shift of the wavelength of light due to the Doppler effect. I.e.,

$$z = \frac{\Delta\lambda}{\lambda_0} = \frac{\lambda - \lambda_0}{\lambda_0} \quad \begin{array}{c} \text{astronomers} \\ \text{notation for} \\ \text{the red shift} \end{array}$$
(13)

where  $\lambda_0$  is the wavelength of the unshifted spectral line, and  $\lambda$  is the Doppler shifted wavelength we see. What is z for galaxy 8C 1435 + 63?

# The Red Shift and the Expanding Universe

In 1917 Albert Einstein published his relativistic theory of gravity, known as General Relativity. In applying his theory of gravity to the behavior of the stars and galaxies in the universe, he encountered what he thought was a serious problem with the theory. Any model of the universe he constructed was unstable. The galaxies tended either to collapse in upon themselves or fly apart. He could not find a solution to his equations that represented the stable unchanging universe everyone knew was out there.

Einstein then discovered that he could add a new term to his gravitational equations. By properly adjusting the value of this term, he could construct a model of the universe that neither collapsed or blew up. This term, that allowed Einstein to create a static model of the universe, became known as the *cosmological constant*.

In later life, Einstein said that his introduction of the cosmological constant was the greatest mistake he ever made. The reason is that the universe is not static. Instead it is *expanding*. The galaxies are all flying apart like the debris from some gigantic explosion. The expansion, or at least instability of the universe, could have been considered one of the predictions of Einstein's theory of gravity, had Einstein not found his cosmological constant. (Later analysis showed that the static model, obtained using the cosmological constant, was not stable. The slightest perturbation would cause it to either expand or contract.)

That the universe is not static was discovered by Doppler shift measurements. In the 1920s, the astronomer Edwin Hubble observed that spectral lines from distant galaxies were all shifted toward the red, and that the farther away the galaxy was, the greater the red shift. Interpreting the red shift as being due to the Doppler effect meant that the distant galaxies were moving away from us, and the farther away a galaxy was, the faster it was moving.

Hubble was the first astronomer to develop a way to measure the distance out to other galaxies. Thus he could compare the red shift or recessional velocity to the distance the galaxy is away from us. He found a simple rule known as Hubble's law. If you look at galaxies twice as far away, they will be receding from us twice as fast. Roughly speaking, he found that a galaxy .1 billion light years away would be receding at 1% the speed of light; a galaxy .2 billion light years away at 2% the speed of light, etc. In the 1930s, construction of the 200 inch Mt. Palomar telescope was started. It was hoped that this telescope (completed in 1946) would be able to observe galaxies as far away as 2 billion light years. Such galaxies should be receding at the enormous speeds of approximately 20% the speed of light.

With the discovery of quasars, we have been able to observe much more distant galaxies, with far greater recessional velocities. As we have just seen, the galaxy 8C 1435+63, photographed by the 10 meter (400 inch) telescope in Hawaii, is receding from us at a speed of 95% the speed of light. To analyze the Doppler effect for such a galaxy, the fully relativistic Doppler effect formula, Equation 12 is needed; non relativistic approximations will not do.

Hubble's law raises several interesting questions. First, it sounds as if we must be at the center of everything, since the galaxies in the universe appear to all be moving away from us. But this is simply a consequence of a uniform expansion. Someone in a distant galaxy will also observe the same Hubble law. To see how a uniform expansion works, mark a number of equally spaced dots on a partially blown up balloon. Select any one of the dots to represent our galaxy, and then start blowing up the balloon to represent the expansion of the universe. You will notice that dots twice as far away move away twice as fast, no matter which dot you selected. Hubble's law is obeyed from the point of view of any of the dots on the balloon. (You can see this expansion in Figure (33), where we started with an array of light colored dots, and uniformly expanded the array to get the black dots.)

Another interesting question is related to nature's speed limit c. We cannot keep looking out twice as far to see galaxies receding twice as fast, because we cannot have galaxies receding faster than the speed of light. Something special has to happen when the recessional speeds approach the speed of light, as they have in the case of 8C 1435 + 63. This appears to place a limit on the size of the universe we can observe.



#### Figure 33

During a uniform expansion, neighboring galaxies move a certain distance away (d), galaxies twice as far away move twice as far (2d), etc. This is Hubble's law for the expanding universe. One of the things to remember when we look at distant galaxies is that we are not only looking far away, but we are also looking back in time. When we look at a galaxy 10 billion light years away, we are looking at light emitted 10 billion years ago, when the universe was 10 billion years younger. Recent studies have clearly shown that galaxies 10 billion years away look different than nearby galaxies. Over the past 10 billion years the universe has evolved; galaxies have aged, becoming more symmetric and less violent.

To predict what we will find as we look back in time, look at ever more distant galaxies, imagine that we take a moving picture of the universe and run the moving picture backwards.

If we reverse the moving picture of expanding galaxies, we see contracting galaxies. They are all contracting back to one point in space and time. Go back to that point and run the movie forward, and we see all of the universe rushing out of that point, apparently the consequence of a gigantic explosion. This explosion has become known as the **Big Bang**. (The name Big Bang was a derisive expression coined by the astronomer Fred Hoyle who had a competing theory of the origin of the universe.)

The idea that the universe started in a big bang, provides a simple picture of the Hubble law. From our point of view, galaxies emerged at various speeds in all directions from the Big Bang. Those that were moving away from us the fastest just after the explosion are now the farthest away from us. Galaxies moving away twice as fast are now twice as far away.

In the next chapter we will have more to say about the origin of the universe and evidence for the Big Bang. We will also introduce another way to interpret the Doppler effect and its relationship to the expansion of the universe.

# A CLOSER LOOK AT INTERFERENCE PATTERNS

Our focus so far in this chapter has been on the application of the wave properties of light to the study of physical phenomena such as atomic spectra and the expansion of the universe. We now wish to turn our attention to a more detailed study of the wave phenomena itself. We will first take a closer look at the single slit diffraction pattern that serves as the envelope of the multiple slit patterns we saw back in Figure (18). We will then discuss an experimental technique for accurately recording various interference patterns produced by laser beams. We will then end the chapter with a demonstration of how Fourier analysis can be used to predict the structure of the interference patterns we observe.

The reason for these studies is to strengthen intuition about the behavior of waves. The remainder of the text deals with the inherent wave nature of matter, and here we wish to develop the conceptual and experimental tools to study this wave nature.

# **Single Slit Diffraction Pattern**

In the 1600s, Francesco Maria Grinaldi discovered that light going through a fine slit cannot be prevented from spreading on the other side. He named this phenomenon *diffraction*. Independently Robert Hook, of Hook's law fame, made the same observation and provided a wave like explanation. The clearest explanation comes from the Huygens construction illustrated in Figures (3) through (8).



**Figure 3 (repeated)** The simple diffraction pattern you get when the slit is narrow compared to a wavelength.

In Figure (3) reproduced here, we have a deceptively simple picture of the single slit diffraction pattern. In the photograph a wave is impinging upon a slit whose width is less than one wavelength, with the result that we get a simple circular wave emerging on the other side. In Figures (6), (7) and (8) we look at what happens when the slit becomes wider than a wavelength. These are all views of the wave pattern close to the slit. We see that as the slit becomes wider, more of the wave passes through undisturbed, creating a more or less distinct shadow effect. Nevertheless we always see circular waves at the edge of the shadow. If you carefully look at Figure (6), reproduced below, you can see lines of nodes coming out of the slit that is about 2 wavelengths wide.

In Figure (34), we have the diffraction pattern produced by a laser beam passing through a 50 micron wide single slit and striking a screen 10 meters away. This is a reproduction of the single slit pattern that acts as an envelope for the multiple slit patterns seen in Figure (18). A 50 micron slit is  $50 \times 10^{-6}$  meters or  $500 \times 10^{-5}$  cm wide. This is nearly a hundred times greater than the  $6.4 \times 10^{-5}$  cm wavelength of the laser light passing through the slit. Thus we are dealing with slits that are about 100 wavelengths wide.



**Figure 6 (repeated)** The pattern becomes more complex when the slit is wider than a wavelength. Here you can begin to see lines of nodes emerging from the slit.

The fact that the diffraction pattern was photographed 10 meters from the slits means that we are looking at the pattern nearly 20 million wavelengths away from the slits. Thus the ripple tank photographs of Figures (5) through (8), showing diffraction patterns within a few wavelengths of the slits, are not a particularly relevant guide as to what we could expect to see 20 million wavelengths away.

The general features of the diffraction pattern in Figure (34) is that we have a relatively broad central maximum, with nodes on either side. Then there are dimmer and narrower maxima on either side. There is a series of these side maxima that extend out beyond the photograph of Figure (34).

If the slit were narrow compared to a wavelength, if the wave spread out as in Figures (3), then we would get just one broad central maxima. Only when the slit is wider than a wavelength do we get the minima we see in Figure (34). These minima result from the interference and cancellation of waves from different parts of the slit. What we wish to do now is to show how this cancellation occurs and predicts where the minima will be located.



For this single slit laser beam diffraction pattern, the slit was about 100 wavelengths wide (w), and the screen was about 20 million wavelengths away (D).

# Analysis of the Single Slit Pattern

In our discussion of diffraction gratings, we estimated the width of the maxima by determining how far from the center of the maxima the intensity first went to zero, where we first got complete cancellation. This occurred where light from pairs of slits cancelled. In our example of Figure (21), light from slit 1 cancelled that from slit 501, from slit 2 with slit 502, etc., all the way down to slits 500 and 1000.

We can use a similar analysis for the single slit pattern, except the one big slit is broken up, *conceptually*, into many narrow slits, as illustrated in Figure (35). Suppose, for example, we think of the one wide slit of width w as being broken up into 1000 neighboring individual slits. The individual slits are so narrow that each piece of wave front in them should act as a source of a pure circular wave as shown back in Figure (3).

Now consider the light heading out in such a direction that the wave from the first conceptual slit is half a wavelength  $\lambda/2$ , in front of the wave from the middle slit, number 501. When the waves from these two "slits" strike the screen they will cancel. Similarly waves from slits 2 and 502 will cancel, as will those from 3 and 503, etc., down to 500 and 1000. Thus is the direction where the path length difference from the edge to the center of the opening is half a wavelength  $\lambda/2$ . Between the two edges of the opening, the path length difference to this minimum is  $\lambda$ , as shown.

Figure 35 Conceptually break the single slit up into many individual slits. We get a minimum when light from the conceptual slits cancels in pairs.



From Figure (36), we can calculate the height of the first minimum using the familiar similar triangles we have seen in previous analysis. The small right triangle near the slit has a short side of length  $\lambda$  and a hypotenuse equal to the slit width w. The big triangle has a short side equal to  $Y_{min}$  and a hypotenuse given by the Pythagorean theorem as  $\sqrt{D^2 + Y_{min}^2}$ . Usually  $Y_{min}$  will be much smaller than the distance D, so that we can replace  $\sqrt{D^2 + Y_{min}^2}$  by D, to get

$$\frac{\lambda}{w} = \frac{Y_{min}}{\sqrt{D^2 + Y_{min}^2}} \approx \frac{Y_{min}}{D}$$

or

$$Y_{\min} \approx \frac{\lambda D}{W} \qquad \begin{array}{l} \text{distance to the} \\ \text{first minima of} \\ a \text{ single slit} \\ \text{diffraction pattern} \end{array}$$
(14)

#### **Exercise 14**

To obtain the single slit diffraction pattern seen in Figure (34), we used a slit 50 microns wide located 10 meters from the screen. The distance  $Y_{min}$  to the first minimum was about 13 cm. Use this result to determine the wavelength of the laser light used. Compare your answer with your results from Exercises 5 and 6, where the same wavelength light was used.



Figure 36

Similar triangles for calculating the distance to the first minimum of a single slit diffraction pattern.

# RECORDING DIFFRACTION GRATING PATTERNS

Another way to record the diffraction pattern is to use a device called a *photoresistor*. A photoresistor is an inexpensive resistor whose resistance  $R_p$  varies depending upon the intensity of the light striking the resistor. If you place the photoresistor in the circuit shown in Figure (37), along with a fixed resistance  $R_2$ , a battery of voltage  $V_b$ , and an oscilloscope, you can measure with the oscilloscope the intensity of light striking the photoresistor.

The analysis of the circuit in Figure (37) is as follows. The resistors  $R_p$  and  $R_1$  are in series and thus have an effective resistance  $R = R_p + R_1$ . The current i in the circuit is thus  $i = V_b/R = V_b/(R_p + R_1)$ . Thus as the photoresistor's resistance  $R_p$  changes with changes in light intensity, the current i will also change. Finally the voltage  $V_1$  that the oscilloscope sees across the fixed resistor  $R_1$  is given by Ohm's law as  $V_1 = iR_1$ . Thus as i changes,  $V_1$  changes and we see the change on the oscilloscope.



#### Figure 37

The photoresistor circuit. By making  $R_2$  considerably bigger than the photoresistor resistance  $R_p$ , the current i stays relatively constant. As a result, the voltage  $V_p = i R_p$  is nearly proportional to  $R_p$ . (We used  $R_p = 6.8K\Omega$  and the EG&G opto VT30N4 photoresistor.) For a number of years we tried various ways of moving the photoresistor through the diffraction pattern in order to record the intensity of the light in the diffraction pattern. We tried mounting the photoresistor on xy recorders and various home-built devices, but there was always some jitter and the results were only fair. The solution, as it turns out, is not to move the photoresistor, but move the diffraction pattern across a fixed photoresistor instead. This is easily done using a rotating mirror, a mirror attached to a clock motor as shown in Figure (38). We have found that if you use a motor with a speed of 1/2 revolutions per minute, you have plenty of time to make a stable noise-free recording of the bottom. Using the recording oscilloscope MacScope, we recorded the single slit diffraction pattern seen in Figure (39).





The rotating mirror. (We were careful to make sure that the axis of rotation was accurately perpendicular to the base. If it isn't, the laser beam wobbles up and down.) A photoresistor is sensitive to the intensity or energy density of the light striking it. And the intensity is proportional to the square of the amplitude of the waves in the beam. Thus in Figure (39) we are looking at a graph of the square of the wave amplitude in a single slit diffraction pattern. It is reasonable that the intensity should be proportional to the square of the amplitude, because amplitudes can be positive or negative, but intensities are always positive. You cannot have a negative intensity, and you do not get one if you square the amplitude since squares of real numbers are always positive.



**Figure 39 a,b** Single slit diffraction pattern. Data from the project by Cham, Cole, and Layang.



#### Figure 39c

Single slit diffraction pattern with the amplitude of the voltage amplified so that we can see the side lobes. Data from the project by Cham, Cole, and Layang.

## **Exercise 15**

In Figure (40) we study the interference pattern produced by a laser beam passing through three equally spaced slits. Figure (40a) shows the experimental setup and (40b) the shape of the slits through which the laser beam went. Figure (40c) is a photograph of the interference pattern, and Figure (40d) is a recording in which the voltage is proportional to the intensity of the light striking a stationary photoresistor. The beam rotated at a rate of 1 revolution per hour, sweeping the beam past the photo resistor. (The mirror only turned at .5 revolutions per hour, but the reflected beam rotates twice as fast as the mirror. You can see this by the fact that when the mirror turns  $45^{\circ}$  the beam rotates  $90^{\circ}$ .)

To Calculate the wavelength of the laser light from the experimental data in Figure 40, first note that the beam is sweeping past the photoresistor at a speed

$$v_{\text{beam}} = \frac{2\pi \, \text{r} \, (\text{cm} / \text{revolution})}{3600 \, (\text{sec} / \text{revolution})} = \frac{2\pi \, \text{r}}{3600} \frac{\text{cm}}{\text{sec}}$$

where r is the distance from the axis of the mirror to the photoresistor. If it takes a time T for two maxima to sweep past the photoresistor, then the distance  $Y_{max}$  between the maxima is

$$Y_{max} = V_{beam} \times T = \frac{2\pi r}{3600} \times T$$

If the slits are close to the mirror, then r is also equal to the distance D from the slits to the photoresistor (screen). The wavelength is then given by Equation 3a as

$$\lambda = Y_{max} \times \frac{d}{D} \approx Y_{max} \times \frac{d}{r}$$
$$= \left(\frac{2\pi r}{3600} \times T\right) \times \frac{d}{r}$$

The factors of r cancel, and we are left with

$$\lambda = \frac{2\pi \,\mathrm{d}}{3600} \times T \tag{15}$$

Thus if the slits are close to the mirror, we do not need to know the distance to the photoresistor. (You can see that if the rotating beam is twice as long, the end travels twice as fast. But the maxima are twice as far apart, thus it takes the same length of time for the maxima to pass the photoresistor.)

Use the results of Figure 40d to determine the wavelength of the laser beam. (In this experiment, the slits were close to the mirror.)

(More to come on the use of Fourier analysis to predict diffraction patterns.)





b) The slits.  $(1\mu = 10^{-6}m = 10^{-4}cm)$ 



c) 3 slit diffraction pattern



d) Voltage recording on photoresistor



# Chapter 34 Photons

The effort to determine the true nature of light has been a fitful process in the history of physics. Newton and Huygens did not agree on whether light was a wave or consisted of beams of particles. That issue was apparently settled by Thomas Young's two-slit experiment performed in 1801, nearly three quarters of a century after Newton's death. Young's experiment still did not indicate what light was a wave of. That insight had to come from Maxwell's theory of 1860 which showed that light was a wave of electric and magnetic fields.

In the late 1800s there were dramatic confirmations of Maxwell's theory. In 1888 Heinrich Hertz observed radio waves, the expected low frequency component of the electromagnetic spectrum. As we have seen from our own experiments, the electric and magnetic fields in a radio wave can be measured directly.

But as the nineteenth century was ending, not all predictions of Maxwell's theory were as successful. Applications of Maxwell's equations to explain the light radiated by matter were not working well. No one understood why a heated gas emitted sharp spectral lines, and scientists like Boltzman were unable to explain important features of light radiated by hot solid objects. The fact that Boltzman could get some features right, but not others, made the problem more vexing. Even harder to understand was the way beams of light could eject electrons from the surface of a piece of metal, a phenomenon discovered in 1897 by Hertz. Many of these problems were cleared up by a picture developed by Max Planck and Einstein, a picture in which light consisted of beams of particles which became known as photons. The photon picture immediately explained the ejection of electrons from a metal surface and the spectrum of radiation from a heated solid object. In the past few years the observation of photons coming in uniformly from all directions in space has led to a new and surprisingly well confirmed picture of the origin of the universe.

In this chapter we will discuss the properties of photons and how discovering the particle nature of light solved some outstanding problems of the late nineteenth century. We will finish with a discussion of what photons have told us about the early universe.

What we will not discuss in this chapter is how to reconcile the two points of view about light. How could light behave as a wave in Thomas Young's experiment, and as a particle in experiments explained by Einstein. How could Maxwell's theory work so well in some cases and fail completely in others? These questions, which puzzled physicists for over a quarter of a century, will be the topic of discussion in the chapter on quantum mechanics.



# **BLACKBODY RADIATION**

When we studied the spectrum of hydrogen, we saw that heated hydrogen gas emits definite spectral lines, the red hydrogen  $\alpha$ , the blue hydrogen  $\beta$  and violet hydrogen  $\gamma$ . Other gases emit definite but different spectral lines. But when we look through a diffraction grating at the heated tungsten filament of a light bulb, we see something quite different. Instead of sharp spectral lines we see a continuous rainbow of all the colors of the visible spectrum. Another difference is that the color of the light emitted by the filament changes as you change the temperature of the filament. If you turn on the light bulb slowly, you first see a dull red, then a brighter red, and finally the filament becomes white hot, emitting the full spectrum seen in white light. In contrast, if you heat hydrogen gas, you see either no light, or you see all three spectral lines at definite unchanging wave lengths.

Some complications have to be dealt with when studying light from solid objects. The heated burner on an electric stove and a ripe McIntosh apple both look red, but for obviously different reasons. The skin of the McIntosh apple absorbs all frequencies of visible light except red, which it reflects. A stove burner, when it is cool, looks black because it absorbs all wavelengths of light equally. When the black stove burner is heated, the spectrum of light is not complicated by selective absorption or emission properties of the surface that might enhance the radiation at some frequencies. The light emitted by a heated black object has universal characteristic properties that do not depend upon what kind of black substance is doing the radiating. The light from such objects is *blackbody radiation*.

One reason for studying blackbody radiation is that you can determine the temperature of an object from the light it emits. For example, Figure (1) shows the intensity of light radiated at different wavelengths by a tungsten filament at a temperature of 5800 kelvins. The greatest intensity is at a wavelength of  $5 \times 10^{-5}$  cm, the middle of the visible spectrum at the color yellow. If we plot intensities of the various wavelengths radiated by the sun, you get essentially the same curve. As a result we can conclude that the temperature of the surface of the sun is 5800 kelvins. It would be hard to make this measurement any other way.

There are a few simple rules governing blackbody radiation. One is that the wavelength of the most intense radiation, indicated by  $\lambda_{max}$  in Figure (1), is inversely proportional to the temperature. The explicit formula, known as *Wein's displacement law* turns out to be

$$\lambda_{\max} = \frac{2.898}{T} \text{mmK} \tag{1}$$

where  $\lambda_{max}$  is in millimeters and the temperature T is in kelvins. For T= 5800K, Equation 1 gives

$$\lambda_{\text{max}}(5800\text{K}) = \frac{2.898\text{mmK}}{5800\text{K}} = 5 \times 10^{-4}\text{mm}$$
  
=  $5.0 \times 10^{-5}$ cm

which is the expected result.

While  $\lambda_{max}$  changes with temperature, the relative shape of the spectrum of radiated intensities does not. Figure (1) is a general sketch of the blackbody radiation spectrum. To determine the blackbody spectrum for another temperature, first calculate the new value of  $\lambda_{max}$  using Equation 1 then shift the horizontal scale in Figure (1) so that  $\lambda_{max}$  has this new value.



## Figure 1

Blackbody spectrum at 5800 degrees on the kelvin scale. The solid line is the experimental curve, the dotted line represents the prediction of Newtonian mechanics combined with Maxwell's equations. The classical theory agrees with the experimental curve only at long wavelengths.
Knowledge of the blackbody spectrum is particularly useful in astronomy. Most stars radiate a blackbody spectrum of radiation. Thus a measurement of the value of  $\lambda_{max}$  determines the temperature of the surface of the star. There happens to be quite a variation in the surface temperature and color of stars. This may seem surprising at first, because most stars look white. But this is due to the fact that our eyes are not color sensitive in dim light. The variation in the color of the stars can show up much better in a color photograph.

As an example of the use of Equation 1, suppose you observe a red star that is radiating a blackbody spectrum with  $\lambda_{max} = 7.0 \times 10^{-5}$  cm. The surface temperature should then be given by

$$T = \frac{2.898 \text{ mmK}}{7.0 \times 10^{-4} \text{ mm}} = 4140 \text{ kelvin}$$

#### Exercise 1

(a) What is the surface temperature of a blue star whose most intense wavelength is  $\lambda_{max} = 4 \times 10^{-5}$  cm?

(b) What is the wavelength  $\lambda_{max}$  of the most intense radiation emitted by an electric stove burner that is at a temperature of 600° C (873K)?

Another feature of blackbody radiation is that the intensity of the radiation increases rapidly with temperature. You see this when you turn up the voltage on the filament of a light bulb. Not only does the color change from red to white, the bulb also becomes much brighter.

The net amount of radiation you get from a hot object is the difference between the amount of radiation emitted and the amount absorbed from the surroundings. If the object is at the same temperature as its surroundings, it absorbs just as much radiation as it emits, with the result that there is no net radiation. This is why you cannot feel any heat from an electric stove burner before it is turned on. But after the burner is turned on and its temperature rises above the room temperature, you begin to feel heat. Even if you do not touch the burner you feel infrared radiation which is being emitted faster than it is being absorbed. By the time the burner becomes red hot, the amount of radiation it emits greatly exceeds the amount being absorbed.

In 1879, Joseph Stefan discovered that the total intensity, the total energy emitted per second in blackbody radiation was proportional to the fourth power of the temperature, to  $T^4$  where T is in kelvins. Five years later Ludwig Boltzman explained the result theoretically. This result is thus known as the Stefan-Boltzman law.

As an example of the use of the Stefan-Boltzman law, suppose that two stars are of the same size, the same surface area, but one is a red star at a temperature of 4,000K while the other is a blue star at a temperature of 10,000K. How much more rapidly is the hot blue star radiating energy than the cool red star?

The ratio of the rates of energy radiation is equal to the ratio of the fourth power of the temperatures. Thus

energy radiated  
by blue star  
energy radiated = 
$$\frac{T_{blue}^4}{T_{red}^4} = \left(\frac{10,000 \text{ K}}{4,000 \text{ K}}\right)^4$$
  
by red star  
=  $2.5^4 \approx 40$ 

We see that the blue star must be burning its nuclear fuel 40 times faster than the red star.

# Planck Blackbody Radiation Law

Boltzman used a combination of Maxwell's equations, Newtonian mechanics, and the theory of statistics to show that the intensity of blackbody radiation increased as the fourth power of intensity. But neither he nor anyone else was able to derive the blackbody radiation spectrum shown in Figure (1). There was some success in predicting the long wavelength side of the curve, but no one could explain why the intensity curve dropped off again at short wavelengths.

In 1900 Max Planck tried a different approach. He first found an empirical formula for a curve that matched the blackbody spectrum. Then he searched for a derivation that would lead to his formula. The idea was to see if the laws of physics, as they were then known, could be modified in some way to explain his empirical blackbody radiation curve. Planck succeeded in the following way. According to Maxwell's theory of light, the amount of radiation emitted or absorbed by a charged particle was related to the acceleration of the particle, and that could vary continuously. Planck found that he could get his empirical formula if he assumed that the electrons in a solid emitted or absorbed radiation only in discrete packets. The energy in each packet had to be proportional to the frequency of the radiation being emitted and absorbed. Planck wrote the formula for the energy of the packets in the form

$$E = hf$$
(2)

where f is the frequency of the radiation. The proportionality constant h became known as *Planck's constant*.

For over two decades physicists had suspected that something was wrong either with Newtonian mechanics, Maxwell's equations, or both. Maxwell was unable to derive a formula that explained the specific heat of gases (except the monatomic noble gases), and no one had the slightest idea why heated gases emitted sharp spectral lines. Planck's derivation of the blackbody radiation formula was the first successful derivation of a phenomena that could not be explained by Newtonian mechanics and Maxwell's equations.

But what did it mean that radiation could be emitted or absorbed only in discrete packets or *quanta* as Planck called them? What peculiar mechanism lead to this *quantization* of the emission and absorption process? Planck did not know.

# THE PHOTOELECTRIC EFFECT

1905 was the year in which Einstein cleared up several outstanding problems in physics. We have seen how his focus on the basic idea of the principle of relativity lead to his theory of special relativity and a new understanding of the structure of space and time. Another clear picture allowed Einstein to explain why light was emitted and absorbed in discrete quanta in blackbody radiation. The same idea also explained a process called the *photoelectric effect*, a phenomenon first encountered in 1887 by Heinrich Hertz.

In the photoelectric effect, a beam of light ejects electrons from the surface of a piece of metal. This phenomenon can be easily demonstrated in a lecture, using the kind of equipment that was available to Hertz. You start with a gold leaf electrometer like that shown in Figure (2), an old but effective device for measuring the presence of electric charge. (This is the apparatus we used in our initial discussion of capacitors.) If a charged object is placed upon the platform at the top of the electrometer, some of the charge will flow down to the gold leaves that are protected from air currents by a glass sided container. The gold leaves, each receiving the same sign of charge, repel each other and spread apart as shown. Very small amounts of charge can be detected by the spreading of the gold leaves. To perform the photoelectric effect experiment, clean the surface of a piece of zinc metal by scrubbing it with steel wool, and charge the zinc with a negative charge. We can be sure that the charge is negative by going back to Ben Franklin's definition. If you rub a rubber rod with cat fur, a negative charge will remain on the rubber rod. Then touch the rubber rod to the piece of zinc, and the zinc will become negatively charged. The presence of charge will be detected by the spreading of the gold leaves.

Now shine a beam of light at the charged piece of zinc. For a source of light use a carbon arc that is generated when an electric current jumps the narrow gap between two carbon electrodes. The arc is so bright that you do not need to use a lens to focus the light on the zinc. The setup is shown in Figure (3).

When the light is shining on the zinc, the gold leaves start to fall toward each other. Shut off or block the light and the leaves stop falling. You can turn on and off the light several times and observe that the gold leaves fall only when the light is shining on the zinc. Clearly it is the light from the carbon arc that is discharging the zinc.



**Figure 3** *Photoelectric effect experiment.* 

carbon arc light source



**Figure 2** The gold leaf electrometer. This is the same apparatus we used back in Figure 26-28 in our study of capacitors.

A simple extension to the experiment is to see what happens if the zinc is given a positive charge. Following Ben Franklin's prescription, we can obtain a positive charge by rubbing a glass rod with a silk cloth. Then touch the positively charged glass rod to the zinc and again you see the gold leaves separate indicating the presence of charge. Now shine the light from the carbon arc on the zinc and nothing happens. The leaves stay spread apart, and the zinc is not discharged by the light.

When we charge the zinc with a negative charge, we are placing an excess of electrons on the zinc. From Gauss's law we know that there cannot be any net charge inside a conductor, thus the excess negative charge, the extra electrons, must be residing in the surface of the metal. The light from the carbon arc, which discharges the zinc, must therefore be knocking these extra electrons out of the metal surface. When we charged the zinc positively, we created a deficiency of electrons in the surface, and no electrons were knocked out.

In the context of Maxwell's equations, it is not particularly surprising that a beam of light should be able to knock electrons out of the surface of a piece of metal. According to Maxwell's theory, light consists of a wave of electric and magnetic fields. An electron, residing on the surface of the zinc, should experience an oscillating electric force when the light shines on the zinc. The frequency of oscillation should be equal to the frequency of the light wave, and the strength of the electric field should be directly related to the intensity of the light. (We saw earlier that the intensity of the light should be proportional to the square of the magnitude of the electric field.) The question is whether the electric force is capable of ejecting an electron from the metal surface. A certain amount of energy is required to do this. For example, in our electron gun experiment we had to heat the filament in order to get an electron beam. It was the thermal energy that allowed electrons to escape from the filament. We now want to know whether the oscillating electric force of the light wave can supply enough energy to an electron for the electron to escape.

There are two obvious conclusions we should reach. One is that we do not want the frequency of oscillation to be too high, because the direction of the electric field reverses on each half cycle of the oscillation. The electron is pushed one way, and then back again. The longer the time it is pushed in one direction, the lower the frequency of the oscillation, the more time the electron has to pick up speed and gain kinetic energy. If the frequency is too high, just as the electron starts to move one way, it is pushed back the other way, and it does not have time to gain much kinetic energy.

The second obvious conclusion is that we have a better chance of ejecting electrons if we use a more intense beam of light. With a more intense beam, we have a stronger electric field which should exert a stronger force on the electron, producing a greater acceleration and giving the electron more kinetic energy. An intense enough beam might supply enough kinetic energy for the electrons to escape.

In summary, we expect that light might be able to eject electrons from the surface of a piece of metal if we use a low enough frequency and an intense enough beam of light. An intense beam of red light should give the best results.

These predictions, based on Maxwell's equations and Newtonian mechanics, are completely wrong!

Let us return to our photoelectric effect demonstration. During a lecture, a student suggested that we make the light from the carbon arc more intense by using a magnifying glass to focus more of the arc light onto the zinc. The more intense beam of light should discharge the zinc faster.

When you use a magnifying glass, you can make the light striking the zinc look brighter. But something surprising happens. The zinc stops discharging. The gold leaves stop falling. Remove the magnifying glass and the leaves start to fall again. *The magnifying glass prevents the discharge.* 

You do not have to use a magnifying glass to stop the discharge. A pane of window glass will do just as well. Insert the window glass and the discharge stops. Remove it, and the gold leaves start to fall again.

How could the window glass stop the discharge? The window glass appears to have no effect on the light striking the zinc. The light appears just as bright. It was brighter when we used the magnifying glass, but still no electrons were ejected. The prediction from Maxwell's theory that we should use a more intense beam of light does not work for this experiment.

What the window glass does is block *ultraviolet radiation*. It is ultraviolet radiation that tans your skin (and can lead to skin cancer). It is difficult to get a tan indoors from sunlight that has gone through a window because the glass has blocked the ultraviolet component of the sun's radiation. Similarly the pane of window glass, or the glass in the magnifying lens, used in the photoelectric effect experiment, prevents ultraviolet radiation from the carbon arc from reaching the zinc. It is the high frequency ultraviolet radiation that is ejecting electrons from the zinc, not the lower frequency visible light. This is in direct contradiction to the prediction of Maxwell's theory and Newton's laws.

Einstein's explanation of the photoelectric effect is simple. He assumed that Newton was right after all, in that light actually consisted of beams of particles. The photoelectric effect occurred when a particle of light, a *photon*, struck an electron in the surface of the metal. All the energy of the photon would be completely absorbed by the electron. If this were enough energy the electron could escape, if not, it could not.

The idea that light actually consisted of particles explains why Planck had to assume that in blackbody radiation, light could only be emitted or absorbed in quantum units. What was happening in blackbody radiation, photons, particles of light, were being emitted or absorbed. As a result, Planck's formula for the energy of the quanta of emitted and absorbed radiation, must also be the formula for the energy of a photon. Thus Einstein concluded that a photon's energy is given by the equation

$$E_{\text{photon}} = \text{hf} \qquad \begin{array}{c} \text{Einstein's photoelectric} \\ \text{effect formula} \end{array} (3)$$

where again f is the frequency of the light and h is Planck's constant. Equation 3 is known as Einstein's *photoelectric effect formula*.

With Equation 3, we can begin to understand our photoelectric effect demonstration. It turns out that visible photons do not have enough energy to knock an electron out of the surface of zinc. There are other metals that require less energy and for these metals visible light will produce a photoelectric effect. But for zinc, visible photons do not have enough energy. Even making the visible light more intense using a magnifying glass does not help. It is only the higher frequency, more energetic, ultraviolet photons that have enough energy to kick an electron out of the surface of zinc. We blocked these energetic photons with the window glass and the magnifying glass.

In 1921, Einstein received the Nobel prize, not for the special theory of relativity which was still controversial, nor for general relativity, but for his explanation of the photoelectric effect.

# **PLANCK'S CONSTANT** h

Planck's constant h, the proportionality constant in Einstein's photoelectric effect formula, appears nowhere in Newtonian mechanics or Maxwell's theory of electricity and magnetism. As physicists were to discover in the early part of the twentieth century, Planck's constant appears just when Newtonian mechanics and Maxwell's equations began to fail. Something was wrong with the nineteenth century physics, and Planck's constant seemed to be a sign of this failure.

The value of Planck's constant is

$$h = 6.63 \times 10^{-34}$$
 joule sec (4)

where the dimensions of h have to be an energy times a time, as we can see from the photoelectric formula

$$E_{joules} = h(joulesec) \times f(\frac{cycles}{sec})$$
 (3a)

The dimensions check because cycles are dimensionless. It is not hard to see that Planck's constant also has the dimensions of angular momentum. Recall that the angular momentum L of an object is equal to the object's linear momentum p = mv times its lever arm  $r_{\perp}$  about some point. Thus the formula for angular momentum is

$$L = pr_{\perp} = m(kg)v(\frac{meter}{sec}) \times r_{\perp}(meter)$$
$$= mvr_{\perp}(kg\frac{m^2}{sec})$$

We get the same dimensions if we write Planck's constant in the form

$$h(\text{joule sec}) = h\left(kg\frac{m^2}{\sec^2}\right) \times \sec$$
$$= h\left(kg\frac{m^2}{\sec}\right)$$
(5)

where we used the fact that the dimensions of energy are a mass times a velocity squared.

A fundamental constant of nature with the dimensions of angular momentum is not something to be expected in Newtonian mechanics. It suggests that there is something special about this amount of angular momentum,  $6.63 \times 10^{-34}$  (kg m<sup>2</sup>/sec) of it, and nowhere in Newtonian mechanics is there any reason for any special amount. It would be Neils Bohr in 1913 who first appreciated the significance of this amount of angular momentum.

# **PHOTON ENERGIES**

Up to a point we have been describing the electromagnetic spectrum in terms of the frequency or the wavelength of the light. Now with Einstein's photoelectric formula, we can also describe the radiation in terms of the energy of the photons in the radiation. This can be convenient, for we often want to know how much energy photons have. For example, do the photons in a particular beam of light have enough energy to kick an electron out of the surface of a given piece of metal, or to break a certain chemical bond?

For visible light and nearby infrared light, the frequencies are so high that describing the light in terms of frequency is not particularly convenient. We are more likely to work in terms of the light's wavelength and the photon's energy, and want to go back and forth between the two. Using the formula

$$f \frac{\text{cycles}}{\text{sec}} = \frac{\text{c meters/sec}}{\lambda \text{ meters/cycle}}$$

which we can get from dimensions, we can write the photoelectric formula in the form

$$E = hf = \frac{hc}{\lambda}$$
(6)

Using MKS units in Equation 6 for h, c, and  $\lambda$ , we end up with the photon energy expressed in joules. But a joule, a huge unit of energy compared to the energy of a visible photon, is also inconvenient to use. A far more convenient unit is the electron volt. To see why, let us calculate the energy of the photons in the red hydrogen  $\alpha$  line, whose wavelength was  $6.56 \times 10^{-5}$ cm or  $6.56 \times 10^{-7}$ m. First calculating the energy in joules, we have

$$E(H_{\alpha \text{ line}}) = \frac{hc}{\lambda_{\alpha}}$$
  
= 
$$\frac{6.63 \times 10^{-34} \text{ joule sec } \times 3 \times 10^8 \text{ m / sec}}{6.56^{-7} \times 10^{-7} \text{ m}}$$
  
= 
$$3.03 \times 10^{-19} \text{ joules}$$

Converting this to electron volts, we get

$$E(H_{\alpha \text{ line}}) = \frac{3.03 \times 10^{-19} \text{ joules}}{1.6 \times 10^{-19} \text{ joules/eV}}$$
$$E(H_{\alpha \text{ line}}) = 1.89 \text{ eV}$$
(7)

That is a convenient result. It turns out that the visible spectrum ranges from about 1.8 eV for the long wavelength red light to about 3.1 eV for the shortest wavelength blue photons we can see. It requires 3.1 eV to remove an electron from the surface of zinc. You can see immediately that visible photons do not quite have enough energy. You need ultraviolet photons with an energy greater than 3.1 eV.

#### **Exercise 2**

The blackbody spectrum of the sun corresponds to an object whose temperature is 5800 kelvin. The predominant wavelength  $\lambda_{max}$  for this temperature is  $5.0\times10^{-5} \rm cm$  as we saw in the calculation following Equation 1. What is the energy, in electron volts, of the photons of this wavelength?

#### **Exercise 3**

The rest energy of an electron is  $.51 \text{MeV} = 5.1 \times 10^5 \text{ eV}$ . What is the wavelength, in centimeters, of a photon whose energy is equal to the rest energy of an electron? We will often want to convert directly from a photon's wavelength  $\lambda$  in centimeters to its energy E in electron volts. This is most easily done by starting with the formula  $E = hc/\lambda$  and using conversion factors until E is in electron volts when  $\lambda$  is in centimeters. We get

$$E = \frac{hc}{\lambda}$$
  
=  $\frac{6.63 \times 10^{-34} \text{ joule sec} \times 3 \times 10^{10} \frac{\text{cm}}{\text{sec}}}{\lambda \text{ cm}}$   
=  $\frac{1.989}{\lambda \text{ cm}} \times 10^{-23} \text{ joule cm} \times \frac{1}{1.6 \times 10^{-19} \frac{\text{joule}}{\text{eV}}}$ 

The desired formula is thus

$$E_{photon}(in eV) = \frac{12.4 \times 10^{-5} eV \cdot cm}{\lambda(in cm)}$$
(8)

As an example in the use of Equation 8, let us recalculate the energy of the  $H_{\alpha}$  photons whose wavelength is  $6.56 \times 10^{-5}$  cm. We get immediately

$$E_{H_{\alpha}} = \frac{12.4 \times 10^{-5} \text{eV cm}}{6.56 \times 10^{-5} \text{cm}} = 1.89 \text{eV}$$

which is our previous result.

#### **Exercise 4**

The range of wavelengths of light in the visible spectrum is from  $7 \times 10^{-5}$  cm in the red down to  $4 \times 10^{-5}$  cm in the blue. What is the corresponding range of photon energies?

#### **Exercise 5**

(a) It requires 2.20 eV to eject an electron from the surface of potassium. What is the longest wavelength light that can eject electrons from potassium?

(b) You shine blue light of wavelength  $4 \times 10^{-5}$  cm at potassium. What is the maximum kinetic energy of the ejected electrons?

#### **Exercise 6**

The human skin radiates blackbody radiation corresponding to a temperature of 32°C. (Skin temperature is slightly lower than the 37°C internal temperature.) What is the predominant energy, in eV of the photons radiated by a human? (This is the energy corresponding to  $\lambda_{max}$  for this temperature.)

#### Exercise 7

A 100 watt bulb uses 100 joules of energy per second. For this problem, assume that all this energy went into emitting yellow photons at a wavelength of  $\lambda = 5.88 \times 10^{-5}$  cm.

(a) What is the energy, in eV and joules, of one of these photons?

(b) How many of these photons would the bulb radiate in one second?

(c) From the results of part (b), explain why it is difficult to detect individual photons in a beam of light.

#### Exercise 8

Radio station WBZ in Boston broadcasts at a frequency of 1050 kilocycles at a power of 50,000 watts.

(a) How many photons per second does this radio station emit?

(b) Should these photons be hard to detect individually?

#### Exercise 9

In what part of the electromagnetic spectrum will photons of the following energies be found?

(a) 1 eV	(e) 5 eV
(b) 2.1 eV	(f) 1000 eV
(c) 2.5 eV	(g) $.51 \times 10^6 eV$ (.51 MeV)
(d) 3 eV	(h) $4.34 \times 10^{-9} eV$

(The rest energy of the electron is .51 MeV.)

#### **Exercise 10**

(a) Calculate the energy, in eV, of the photons in the three visible spectral lines in hydrogen

$$\begin{split} \lambda_{\alpha}(\text{red}) &= 6.56 \times 10^{-5} \text{cm} \\ \lambda_{\beta}(\text{blue}) &= 4.86 \times 10^{-5} \text{cm} \\ \lambda_{\gamma}(\text{violet}) &= 4.34 \times 10^{-5} \text{cm} \end{split}$$

It requires 2.28 eV to eject electrons from sodium.

(b) The red  $H_{\alpha}$  light does not eject electrons from sodium. Explain why.

(c) The H\_{\beta} and H\_{\gamma} lines do eject electrons. What is the maximum kinetic energy of the ejected electrons for these two spectral lines?

# **PARTICLES AND WAVES**

We gain two different perspectives when we think of the electromagnetic spectrum in terms of wavelengths and in terms of photon energies. The wavelength picture brings to mind Young's two slit experiment and Maxwell's theory of electromagnetic radiation. In the photon picture we think of electrons being knocked out of metals and chemical bonds being broken. These pictures are so different that it seems nearly impossible to reconcile them. Reconciling these two pictures will, in fact, be the main focus of the remainder of the text.

For now we seek to answer a more modest question. How can the two pictures coexist? How could some experiments, like our demonstration of the photoelectric effect exhibit only the particle nature and completely violate the predictions of Maxwell's equations, while other experiments, like our measurements of the magnetic field of a radio wave, support Maxwell's equations and give no hint of a particle nature?

In Figure (4) we show the electromagnetic spectrum both in terms of wavelengths and photon energies. It is in the low energy, long wavelength region, from radio waves to light waves, that the wave nature of the radiation tends to dominate. At shorter wavelengths and higher photon energies, from visible light through  $\gamma$  rays, the particle nature tends to dominate. The reason for this was well illustrated in Exercise 8.

In Exercise 8 you were asked to calculate how many photons were radiated per second by radio station WBZ in Boston. The station radiates 50,000 watts of power at a frequency of 1.05 megacycles. To solve the

problem, you first had to calculate the energy of a 1.05 megacycle photon using Einstein's formula  $E_{photon} = hf$ . This turns out to be about  $7 \times 10^{-28}$  joules. The radio station is radiating 50,000 joules of energy every second, and thus emitting  $7 \times 10^{-31}$  photons per second. It is hard to imagine an experiment in which we can detect individual photons when so many are being radiated at once. Any experiments should detect some kind of average effect, and that average effect is given by Maxwell's equations.

When we get up to visible photons, whose energies are in the 2—3 eV range and wavelengths of the order of  $5 \times 10^{-5}$  cm, it is reasonably easy to find experiments that can detect either the particle or the wave nature of light. With a diffraction grating we have no problem measuring wavelengths in the range of  $10^{-5}$  cm. With the photoelectric effect, we can easily detect individual photons in the 2-3 eV range.

As we go to shorter wavelengths, individual photons have more energy and the particle nature begins to dominate. To detect the wave nature of X rays, we need something like a diffraction grating with line spacing of the order of the X ray wavelength. It turns out that the regular lines and planes of atoms in crystalline materials act as diffraction gratings allowing us to observe the wave nature of X ray photons. But when we get up into the  $\gamma$  ray region, where photons have energies comparable to the rest energies of electrons and protons, all we observe experimentally are particle reactions. At these high energies, the wave nature of the photon is basically a theoretical concept used to understand the particle reactions.



While it is a rule of thumb that at wavelengths longer than visible light, the wave nature of electromagnetic radiation dominates, there are important exceptions. The individual photons in the WBZ radio wave can be detected! You might ask, what kind of experiment can detect an object whose energy is only  $7 \times 10^{-28}$ joules. This, however, happens to be the amount of energy required to flip the spin of an electron or a nucleus in a reasonably sized magnetic field. This spin flip process for electrons is called *electron spin reso*nance and for nuclei, nuclear spin resonance. In Chapter 38 we will discuss an electron spin resonance experiment that is easily performed in the lab. Nuclear spin resonance, as you may be aware, is the basis of magnetic resonance imaging, an increasingly important medical diagnostic tool.

The truly amazing feature of the magnetic resonance experiments is that Maxwell's equations and Einstein's photoelectric effect formula make the same predictions! Einstein's photoelectric effect formula is easier to use and will be the way we analyze the electron spin resonance experiment. Maxwell's equations and the classical pictures of angular momentum and gyroscopes facilitate the more detailed analysis needed for the imaging apparatus. Texts describing the imaging apparatus use the classical approach. For this discussion the important point is that the two points of view come together in this low energy, long wavelength limit.

#### **PHOTON MASS**

The basic idea behind Einstein's famous formula  $E = mc^2$  is that energy is mass. The factor  $c^2$  is a conversion factor to go between energy measured in grams and energy measured in ergs. If we had used a different set of units, for example, measuring distances in feet, and time in nanoseconds, then the numerical value of c would be 1, and Einstein's equation would be E = m, the more revealing statement.

Photons have energy, thus they have mass. If we combine the photoelectric formula E = hf with  $E = mc^2$ , we can solve for the mass m of a photon of frequency f. The result is

$$E = hf = m_{photon}c^{2}$$

$$m_{photon} = \frac{hf}{c^{2}}$$
(10)

We can also express the photon mass in terms of the wavelength  $\lambda$ , using f/c = 1 /  $\lambda$ 

$$m_{\text{photon}} = \frac{h}{c} \left( \frac{f}{c} \right) = \frac{h}{c\lambda}$$
 (11)

The idea that photons have mass presents a certain problem. In our earliest discussions of mass in Chapter 6, we saw that the mass increased with velocity, increasing without bounds as the speed of the object approached the speed of light. The formula that described this increase in mass was

$$m = \frac{m_0}{\sqrt{1 - v^2/c^2}} \tag{6-14}$$

where  $m_0$  is the mass of the particle at rest and m its mass when traveling at a speed v.

The obvious problems with photons is that they are light—and therefore travel *at* the speed of light. Applying Equation 6-14 to photons gives

$$m_{photon} = \frac{m_0}{\sqrt{1 - c^2/c^2}} = \frac{m_0}{\sqrt{1 - 1}} = \frac{m_0}{0}$$
 (12)

a rather embarrassing result. The divisor in Equation 12 is exactly zero, not approximately zero. Usually division by 0 is a mathematical disaster.

There is only one way Equation 12 can be salvaged. The numerator  $m_0$  must also be identically zero. Then Equation 12 gives m = 0/0, an undefined, but not disastrous result. The numerical value of 0/0 can be anything -1, 5,  $10^{-17}$ , anything you want. In other words if the rest mass  $m_0$  of a photon is zero, Equation 12 says nothing about what the actual mass  $m_{photon}$  is. Equation 12 only tells us that the rest mass of a photon must be zero.

Stop a photon and what do you have left? Heat! In the daytime many billions of photons strike your skin every second. But after they hit nothing is left except the warmth of the sunlight. When a photon is stopped it no longer exists—only its energy is left behind. That is what is remarkable about photons. Only if they are moving *at* the speed of light do they exist, carry energy and have mass. This distinguishes them from all the particles that have rest mass and cannot get up to the speed of light.

An interesting particle is the neutrino. We are not sure whether a neutrino (there are actually 3 different kinds of neutrinos) has a rest mass or not. If neutrinos have no rest mass, then they must travel at the speed of light, and obey the same mechanics as a photon. The evidence is highly suggestive of this interpretation. We saw, for example, that neutrinos from the 1987 supernova explosion raced photons for some 100,000 years, and took within an hour of the same amount of time to get here. That is very close to the speed of light. If the neutrinos took a tiny bit longer to reach us, if they moved at slightly less than the speed of light, then they would have to have some rest mass. The rest mass of an individual neutrino would have to be extremely small, but there are so many neutrinos in the universe that their total mass could make up a significant fraction of the mass in the universe. This might help explain some of the missing mass in the universe that astronomers are worrying about. At the present time, however, all experiments are consistent with the idea that a neutrino's rest mass is exactly zero.

For particles with rest mass, we used the formulas  $E = mc^2$ ,  $m = m_0/\sqrt{1 - v^2/c^2}$  to get the formulas for the rest energy and the kinetic energy of the particle. In particular we got the approximate formula  $1/2 m_0 v^2$  for the kinetic energy of a slowly moving particle. For photons, the formula  $m_0 / \sqrt{1 - v^2/c^2}$  does not apply, there is no such thing as a slowly moving photon, and the kinetic energy formula  $1/2 mv^2$  is *completely wrong*! For photons, all the energy is kinetic energy, and the formula for the photon's kinetic energy is given by Einstein's photoelectric effect formula  $E = hf = hc/\lambda$ . The energy of a photon is determined by its frequency, not its speed.

#### **Photon Momentum**

While photons have no rest mass, and do not obey Newton's second law, they do obey what turns out to be a quite simple set of rules of mechanics. Like their massive counterparts, photons carry energy, linear momentum, and angular momentum all of which are conserved in interactions between particles. The formulas for these quantities can all be obtained straightforwardly from Einstein's photoelectric formula E = hfand energy formula  $E = mc^2$ .

We have already combined these two equations to obtain Equation 11 for the mass of a photon

$$m_{photon} = \frac{h}{\lambda c}$$
 (11a)

To find the momentum of the photon, we multiply its mass by its velocity. Since all photons move at the same speed c, the photon momentum  $p_{photon}$  is given by

$$p_{\text{photon}} = m_{\text{photon}} c = \frac{h}{\lambda}$$
(13)

In the next few chapters, we will find that Equation 13 applies to more than just photons. It turns out to be one of the most important equations in physics.

In our discussion of systems of particles in Chapter 11, we had an exercise where a boy washing a car, was squirting the hose at the door of the car. The water striking the door carried a certain amount of momentum per second, and as a result exerted a force  $\vec{F} = d\vec{p}/dt$  on the door. The exercise was to calculate this force.

When you shine a beam of light at an object, if the photons in the beam actually carry momentum  $p = h/\lambda$  then the beam should exert a force equal to the rate at which momentum is being absorbed by the object. If the object absorbs the photon, like a black surface would, the momentum delivered is just the momentum of the photons. If it is a reflecting surface, then we have to include the photon recoil, and the momentum transferred is twice as great.

There is a common toy called a radiometer that has 4 vane structures balanced on the tip of a needle as shown in Figure (5). One side of each vane is painted black, while the other side is reflecting. If you shine a beam of light at the vanes, they start to rotate. If, however, you look at the apparatus for a while, you will notice that the vanes rotate the wrong way. They move as if the black side were being pushed harder by the beam of light than the reflecting side.

In the toy radiometers, it is not the force exerted by the light, but the fact that there are some air molecules remaining inside the radiometer, that causes the vanes to rotate. When the light strikes the vanes, it heats the



**Figure 5** *The radiometer.* 

black side more than the reflecting side. Air molecules striking the black side are heated, gain thermal energy, and bounce off or recoil from the vane with more speed than molecules bouncing off the cooler reflecting side. It is the extra speed of the recoil of the air molecules from the black side that turns the vane. This thermal effect is stronger than the force exerted by the light beam itself.

We can see from the example of the radiometer that the measurement of the force exerted by a beam of light, measuring the so-called pressure of light, must be done in a good vacuum during a carefully controlled experiment. That measurement was first made by Nichols and Hull at Dartmouth College in 1901. While Maxwell's theory of light also predicts that a beam of light should exert a force, we can now interpret the Nichols and Hull experiment as the first experimental measurement of the momentum carried by photons.

The first experiments to demonstrate that individual photons carried momentum were carried out by Arthur Compton in 1923. In what is now known as the *Compton scattering* or the *Compton effect*, X ray photons are aimed at a thin foil of metal. In many cases the X ray photons collide with and scatter an electron rather than being absorbed as in the photoelectric effect. Both the struck electron and the scattered photon emerge from the back side of the foil as illustrated in Figure (6).

The collision of the photon with the electron in the metal foil is in many ways similar to the collision of the two steel balls studied in Chapter 7, Figures (1) and (2). The energy of the X ray photons used by Compton were of the order of 10,000 eV while the energy of the electron in the metal is of the order of 1 or 2 eV. Thus the X ray photon is essentially striking an electron at rest, much as the moving steel ball struck a steel ball at rest in Figure (7-2).

In both the collision of the steel balls and in the Compton scattering, both energy and linear momentum are conserved. In particular the momentum carried in by the incoming X ray photon is shared between the scattered X ray and the excited electron. This means that the X ray photon loses momentum in the scattering process. Since the photon's momentum is related to its wavelength by  $p = h/\lambda$ , a loss in momentum means an increase in wavelength. Thus, if the photon mechanics we have developed applies to X ray photons, then the scattered X rays should have a slightly longer wavelength than the incident X rays, a result which Compton observed.

According to Maxwell's theory, if a light wave impinges on a metal, it should start the electrons oscillating at the frequency of the incident wave. The oscillating electrons should then radiate light at the same frequency. This radiated light would appear as the scattered light in Compton's experiments. Thus Maxwell's theory predicts that the scattered X rays should have the same wavelength as the incident wave, a result which is not in agreement with experiment.

While the experiments we have just discussed involved delicate measurements in order to detect the photon momentum, in astronomy the momentum of photons and the pressure of light can have dramatic effects. In about 5 billion years our sun will finish burning the hydrogen in its core. The core will then cool and start to collapse. In one of the contradictory features of stellar evolution, the contracting core releases gravitational potential energy at a greater rate than energy was released by burning hydrogen. As a result the core becomes hotter and much brighter than it was before. The core will become so bright, emit so much light, that the pressure of the escaping light will lift the surface of the sun out into space. As a result the sun will expand until it engulfs the orbit of the earth. At this point the sun will have become what astronomers call a*red giant* star. Because of its huge surface area it will become thousands of times brighter than it is now.

The red giant phase does not last long, only a few million years. If the sun were bigger than it is, the released gravitational potential energy would be enough to ignite helium and nuclear fusion would continue. But the red giant phase for the sun will be near the end of the road. The sun will gradually cool and shrink, becoming a white dwarf star about the size of the earth, and finally a black ember of about the same size.

The pressure of light played an even more important role in the evolution of the early universe. The light from the big bang explosion that created the universe was so intense that for the first 1/3 of a million years, it knocked the particles of matter around and prevented the formation of stars, and galaxies. But a dramatic event occurred when the universe reached an age of 1/3 of a million years. That was the point where the universe had cooled enough to become transparent. At that point the light from the big bang decoupled from matter and stars and galaxies began to form. We will discuss this event in more detail shortly.





# ANTIMATTER

The fact that photons have no rest mass and travel only at the speed of light makes them seem quite different from particles like an electron or proton that have rest mass and make up the atoms and molecules. The distinction fades somewhat when we consider a process in which a photon is transformed into two particles with rest mass. The two particles can be any particleantiparticle pair. Figure (7) is a bubble chamber photograph of the creation of an electron-positron pair by a photon.

In 1926 Erwin Schrödinger developed a wave equation to describe the behavior of electrons in atoms. The first equation he tried had a serious problem; it was a relativistic wave equation that appeared to have two solutions. One solution represented the ordinary electrons he was trying to describe, but the other solution appeared to represent a particle with a negative rest mass. Schrödinger found that if he went to the non relativistic limit, and developed an equation that applied only to particles moving at speeds much less than the speed of light, then the negative rest mass solutions did not appear. The non relativistic equation was adequate to describe most chemical phenomena, and is the famous Schrödinger equation.

A year later, Paul Dirac developed another relativistic wave equation for electrons. The equation was specifically designed to avoid the negative mass solutions, but the techniques used did not work. Dirac's equation correctly predicted some important relativistic phenomena, but as Dirac soon found out, the negative mass solutions were still present.

Usually one ignores undesirable solutions to mathematical equations. For example, if you want to solve for the hypotenuse of a triangle, the Pythagorean theorem tells you that  $c^2 = a^2 + b^2$ . This equation has two solutions,  $c = \sqrt{a^2 + b^2}$  and  $c = -\sqrt{a^2 + b^2}$ . Clearly you want the positive solution, the negative solution in this case is irrelevant.

The problem Dirac faced was that he could not ignore the negative mass solution. If he started with a collection of positive mass particles and let them interact, the equation predicted that negative mass particles would appear, would be created. He could not avoid them.

Through a rather incredible trick, Dirac was able to reinterpret the negative mass solutions as positive mass solutions of another kind of matter—antimatter. In this interpretation, every elementary particle has a corresponding antiparticle. The antiparticle had the same rest mass but opposite charge from its corresponding particle. Thus a particle-antiparticle pair could be created or annihilated without violating the law of conservation of electric charge.

In 1927 when Dirac proposed his theory, no one had seen any form of antimatter, and no one was sure of exactly what to look for. The proton had the opposite charge from the electron, but its mass was much greater, and therefore it could not be the electron's antiparticle. If the electron antiparticle existed, it would have to have the same positive charge as the proton, but the same mass as an electron. In 1932 Carl Anderson at Caltech found just such a particle among the cosmic rays that rain down through the earth's atmosphere. That particle is the **positron** which is shown being created in the bubble chamber photograph of Figure (7).

(In the muon lifetime moving picture, discussed in Chapter 1, positively charged muons were stopped in the block of plastic, emitting the first pulse of light. When a positive muon decays, it decays into a positron and a neutrino. It was the positron that made the second flash of light that was used to measure the muon's lifetime.)

In the early 1950s, the synchrotron at Berkeley, the one shown in Figure (28-27b) was built just large enough to create antiprotons, and succeeded in doing so. Since then we have created antineutrons, and have observed antiparticles corresponding to all the known elementary particles. Nature really has two solutions—matter and antimatter. The main question we have now concerning antimatter is why there is so little of it around at the present time. In the very early universe, temperatures were so high that there was a continual creation and annihilation of particles and antiparticles, with roughly equal but not exactly equal, numbers of particles and antiparticles. There probably was an excess of particles over antiparticles in the order of about one part in 10 billion. In a short while the universe cooled to the point where annihilation became more likely than creation, and the particle-antiparticle pairs annihilated. What was left behind was the slight excess of matter particles, the particles that now form the stars and galaxies of the current universe.

In 1964, James Cronin and Val Fitch, while working on particle accelerator experiments, discovered interactions that lead to an excess of particles over antiparticles. It could be that these interactions were active in the very early universe, creating the slight excess of matter over antimatter. But on the other hand, there may not have been time for known processes to create the observed imbalance. We do not yet have a clear picture of how the excess of matter over antimatter came about.



#### Figure 7

Creation of positron-electron pair. A photon enters from the bottom of the chamber and collides with a hydrogen nucleus. The nucleus absorbs some of the photon's momentum, allowing the photon's energy to be converted into a positron-electron pair. Since a photon is uncharged, it leaves no track in the bubble chamber; the photon's path is shown by a dotted line. (Photograph copyright The Ealing Corporation, Cambridge, Mass.)

#### Exercise 11

Since an electron and a positron have opposite charge, they attract each other via the Coulomb electric force. They can go into orbit forming a small atom like object called *positronium*. It is like a hydrogen atom except that the two particles have equal mass and thus move about each other rather than having one particle sit at The positronium atom lasts for about a the center. microsecond, whereupon the positron and electron annihilate each other, giving off their rest mass energy in the form of photons. The rest mass energy of the electron and positron is so much greater than their orbital kinetic energy, that one can assume that the positron and electron were essentially at rest when they annihilated. In the annihilation both momentum and energy are conserved.

(a) Explain why the positron and electron cannot annihilate, forming only one photon. (What conservation law would be violated by a one photon annihilation?)

(b) Suppose the positronium annihilated forming two photons. What must be the energy of each photon in eV? What must be the relative direction of motion of the two photons?

The answer to part (b) is that each photon must have an energy of .51 MeV and the photons must come out in exactly opposite directions. By detecting the emerging photons you can tell precisely where the positronium annihilated. This phenomenon is used in the medical imaging process called *positron emission tomography* or *PET* scans.

# INTERACTION OF PHOTONS AND GRAVITY

Because photons have mass, we should expect that photons should interact with gravity. But we should be careful about applying the laws of Newtonian gravity to photons, because Newtonian gravity is a non relativistic theory, while photons are completely relativistic particles.

If we apply the ideas of Newtonian gravity to photons, which we will do shortly, we will find that we get agreement with experiment if the photons are moving parallel to the gravitational force, for example, falling toward the earth. But if we do a Newtonian type of calculation of the deflection of a photon as it passes a star, we get half the deflection predicted by Einstein's general theory of relativity. It was in Eddington's famous eclipse expedition of 1917 where the full deflection predicted by Einstein's theory was observed. This observation, along with measurements of the precession of Mercury's orbit, were the first experimental evidence that Newton's theory of gravity was not exactly right.

In 1960, R. V. Pound and G.A. Rebka performed an experiment at Harvard that consisted essentially of dropping photons down a well. What they did was to aim a beam of light of precisely known frequency down a vertical shaft about 22 meters long, and observed that the photons at the bottom of the shaft had a slightly higher frequency, i.e., had slightly more energy than when they were emitted at the top of the shaft.

The way you can use Newtonian gravity to explain their results is the following. If you drop a rock of mass m down a shaft of height h, the rock's gravitational potential energy mgh at the top of the shaft is converted to kinetic energy at the bottom. For a rock, the kinetic energy shows up in the form of increased velocity, and is given by the formula  $1/2 \text{ mv}^2$ . For a photon, all of whose energy is kinetic energy anyway, the kinetic energy gained from the fall shows up as an increased frequency of the photon.

Using Einstein's formula E = hf for the kinetic energy of a photon, we predict that the photon energy at the bottom is given by

$$E_{bottom} = hf_{bottom} = hf_{top} + m_{photon}gy$$
 (14)

where we are assuming that the same formula mgy for gravitational potential energy applies to both rocks and photons.

Since  $m_{photon} = hf/c^2$ , the mass of the photon changes slightly as the photon falls. But for a 22 meter deep shaft, the change in frequency is very small and we can quite accurately use  $hf_{top}/c^2$  for the mass of the photon in Equation 14. This gives

$$hf_{bottom} = hf_{top} + \left(\frac{hf_{top}}{c^2}\right)gy$$

Cancelling the h's, we get

$$f_{bottom} = f_{top} \left( 1 + \frac{gy}{c^2} \right)$$
(15)

as the formula for the increase in the frequency of the photon. This is in agreement with the results found by Pound and Rebka.

#### Exercise 12

(a) Show that the quantity  $gy/c^2$  is dimensionless.

(b) What is the percentage increase in the frequency of the photons in the Pound-Rebka experiment? (Answer:  $2.4 \times 10^{-13}$ %. This indicates how extremely precise the experiment had to be.)

To calculate the sideways deflection of a photon passing a star, we could use Newton's second law in the form  $\vec{F} = d\vec{p}/dt$  to calculate the rate at which a sideways gravitational force added a sideways component to the momentum of the photon. The gravitational force would be  $\vec{F}_g = m_{photon}\vec{g}$ , with  $m_{photon} = hf/c^2$ . The result, as we have mentioned, is half the deflection predicted by Einstein's theory of gravity and half that observed during Eddington's eclipse expedition.

The gravitational deflection of photons, while difficult to detect in 1917, has recently become a useful tool in astronomy. In 1961, Allen Sandage at Mt. Palomar Observatory discovered a peculiar kind of object that seemed to be about the size of a star but which emitted radio waves like a radio galaxy. In 1963 Maarten Schmidt photographed the spectral lines of a second radio star and discovered that the spectral lines were all shifted far to the red. If this red shift were caused by the Doppler effect, then the radio star would be moving away from the earth at a speed of 16% the speed of light.

If the motion were due to the expansion of the universe, then the radio star would have to be between one and two billion light years away. An object that far away, and still visible from the earth, would have to be as bright as an entire galaxy.

The problem was the size of the object. The intensity of the radiation emitted by these radio stars was observed to vary significantly over times as short as weeks to months. This virtually guarantees that the object is no bigger than light weeks or light months across, because the information required to coordinate a major change in intensity cannot travel faster than the speed of light. Thus Schmidt had found an object, not much bigger than a star, radiating as much energy as the billions of stars in a galaxy. These rather dramatic objects, many more of which were soon found, became known as *quasars*, which is an abbreviation for *quasi stellar objects*. It was hard to believe that something not much bigger than a star could be as bright as a galaxy. There were suggestions that the red shift detected by Maarten Schmidt was due to something other than the expansion of the universe. Perhaps quasars were close by objects that just happened to be moving away from us at incredible speeds. Perhaps they were very massive objects so massive that the photons escaping from the object lost a lot of their energy and emerged with lower frequencies and longer wavelengths. (This would be the opposite effect than that seen in the Pound-Rebka experiment where photons falling toward the earth gained kinetic energy and increased in frequency.)

Over the years, no explanation other than the expansion of the universe satisfactorily explained the huge red shifts seen in quasars, but there was this nagging doubt about whether the quasars were really that far away. Everything seemed to fit with the model that red shifts were caused by an expanding universe, but it would be nice to have direct proof.

The direct proof was supplied by gravitational lensing, a consequence of the sideways deflection of photons as they pass a massive object. In 1979, a photograph revealed two quasars that were unusually close to each other. Further investigation showed that the two quasars had identical red shifts and emitted identical spectral lines. This was too much of a coincidence. The two quasars had to be two images of the same quasar. How could two images of a single quasar appear side by side on a photographic plate? The answer is illustrated in Figure (8). Suppose the quasar were directly behind a massive galaxy, so that the light from the quasar to the earth is deflected sideways as shown. Here on earth we could see light coming from the quasar from 2 or more different directions. The telescope forms images as if the light came in a straight line. Thus in Figure (8), light that came around the top side of the galaxy would look like it came from a quasar located above the actual quasar, while light that came around the bottom side would look as if it came from another quasar located below the actual quasar.

This gravitational lensing turned out to be a more common phenomena than one might have expected. More than a dozen examples of gravitational lensing have been discovered in the past decade. Figure (9), an image produced by the repaired Hubble telescope, shows a quasar surrounded by four images of itself. The four images were formed by the gravitational lensing of an intermediate galaxy.

The importance of gravitational lensing is that it provides definite proof that the imaged objects are more distant than the objects doing the imaging. The quasar in Figure (9) must be farther away from us than the



galaxy that is deflecting the quasar's light. This proved that the quasars are distant objects and that the red shift is definitely due to the expansion of the universe.

Evidence over the years has indicted that quasars are the cores of newly formed galaxies. Quasars tend to be distant because most galaxies were formed when the universe was relatively young. If all quasars we see are very far away, the light from them has taken a long time to reach us, thus they must have formed a long time ago. The fact that we see very few nearby quasars means that most galaxy formation has already ceased.

Although we have photographed galaxies for over a hundred years, we know surprisingly little about them, especially what is at the core of galaxies. Recent evidence indicates that at the core of the galaxy M87 there is a black hole whose mass is of the order of millions of suns. The formation of such a black hole would produce brilliant radiation from a very small region of space, the kind of intense localized radiation seen in quasars. At this point we only have proof for one black hole at the center of one galaxy, but the pieces are beginning to fit together. Something quite spectacular may be at the center of most galaxies, and quasars are probably giving us a view of the formation of these centers.



**Figure 9** Hubble telescope photograph of a distant quasar surrounded by 4 images of the quasar. This is known as the Einstein cross.

# **EVOLUTION OF THE UNIVERSE**

The two basic physical ideas involved in understanding the early universe are its expansion, and the idea that the universe was in thermal equilibrium. Before we see how these concepts are applied, we wish to develop a slightly different perspective of these two concepts. First we will see how the red shift of light can be interpreted in terms of the expansion of the universe. Then we will see that blackbody radiation can be viewed as a gas of photons in thermal equilibrium. With these two points of view, we can more easily follow the evolution of the universe.

# Red Shift and the Expansion of the Universe

The original clue that we live in an expanding universe was from the red shift of light from distant galaxies. We have explained this red shift as being caused by the Doppler effect. The distant galaxies are moving away from us, and it is the recessional motion that stretches the wavelengths of the radiated waves, as seen in the ripple tank photograph back in Figure (33-29) reproduced here.



**Figure 33-29** *The doppler effect* 

There is another way to view the red shift that gives the same results but provides a more comprehensive picture of the evolution of the universe. Consider a galaxy that is, for example, receding from us at 10% the speed of light. According to the Doppler effect, the wavelength of the light from that galaxy will be lengthened by a factor of 10%.

Where is that galaxy now? If the galaxy were moving away from us at 10% the speed of light, it has traveled away from us 1/10th as far as the light has traveled in reaching us. In other words the galaxy is 10% farther away now than when it emitted the light. If the recessional motion of the galaxy is due to the expansion of the universe, then the universe is now 10% bigger than it was when the galaxy emitted the light.

In this example, the universe is now 10% bigger and the wavelength of the emitted light is 10% longer. We can take the point of view that *the wavelength of the light was stretched 10% by the expansion of the universe*.

In other words it makes no difference whether we say that the red shift was caused by the 10% recessional velocity of the galaxy, or the 10% expansion of the universe. Both arguments give the same answer. When we are studying the evolution of the universe, it is easier to use the idea that the universe's expansion stretches the photon wavelengths. This is especially true for discussions of the early universe where recessional velocities are close to the speed of light and relativistic Doppler calculations would be required.

#### Another View of Blackbody Radiation

The surface of the sun provides an example of a hot gas more or less in thermal equilibrium. Not only are the ordinary particles, the electrons, the protons, and other nuclei in thermal equilibrium, so are the photons, and this is why the sun emits a blackbody spectrum of radiation. Blackbody radiation at a temperature T can be viewed as a gas of photons in thermal equilibrium at that temperature.

In our derivation of the ideal gas law, we were surprisingly successful using the idea that the average gas molecule had a thermal kinetic energy 3/2 kT. In a similar and equally naive derivation, we can explain one of the main features of blackbody radiation from the assumption that the average or typical photon in blackbody radiation also has a kinetic energy 3/2 kT.

The main feature of blackbody radiation, that could not be explained using Maxwell's theory of light, was the fact that there was a peak in the blackbody spectrum. There is a predominant wavelength which we have called  $\lambda_{max}$  that is inversely proportional to the temperature T. The precise relationship given by Wein's displacement law is

$$\lambda_{\text{max}} = \frac{2.898}{\text{T}} \text{mmK}$$
 (1) repeated

a result we stated earlier. The blackbody radiation peaks around  $\lambda_{max}$  as seen in Figure (1) reproduced here.

If blackbody radiation consists of a gas of photons in thermal equilibrium at a temperature T, we can assume that the average photon should have a kinetic energy like 3/2 kT. (The factor 3/2 is not quite right for relativistic particles, but close enough for this discussion.) Some photons should have more energy, some less, but there should be a peak in the distribution of photons around this energy. Using Einstein's photoelectric effect formula we can relate the most likely photon energy to a most likely wavelength  $\lambda_{max}$ . We have

$$E_{\text{photon}} = \frac{3}{2}kT$$
$$E_{\text{photon}} = \frac{hc}{\lambda_{\text{max}}}$$

Combining these equations gives

$$\frac{hc}{\lambda_{max}} = \frac{3}{2}kT$$
$$\lambda_{max} = \frac{2hc}{3kT}$$

Putting in numbers gives

$$\lambda_{\text{max}} = \frac{2 \times 6.63 \times 10^{-34} \text{ joule sec} \times 3 \times 10^8 \text{ } \frac{\text{m}}{\text{s}}}{3 \times 1.38 \times 10^{-23} \text{ joule}} \text{ } \text{K} \text{ } \text{T}}$$
$$= \frac{.0096 \text{ meter} \cdot \text{K}}{\text{T}}$$

Converting from meters to millimeters gives

$$\lambda_{\max} = \frac{9.60 \text{ mm} \cdot \text{K}}{\text{T}} \quad our \text{ estimate} \qquad (16)$$

While this is not the exact result, it gives us the picture that there should be a peak in the blackbody spectrum around  $\lambda_{max}$ . The formula gives the correct temperature dependence, and the constant is only off by a factor of 3.3. Not too bad a result considering that we did not deal with relativistic effects and the distribution of energies in thermal equilibrium. None of these results can be understood without the photon picture of light.





Blackbody radiation spectrum showing the peak at  $\lambda_{max}$ . (The classical curve goes up to infinity at  $\lambda = 0$ .)

# **MODELS OF THE UNIVERSE**

As we saw in Chapter 33, Hubble was able to combine his new distance scale for stars and galaxies with Doppler shift measurements to discover that the universe is expanding, that the farther a galaxy is away from us, the faster it is moving away from us. Another property of the interaction of light with matter, the blackbody spectrum discussed at the beginning of this chapter, provided a critical clue to the role of this expansion in the history of the universe. To see why, it is instructive to look at the evolution of our picture of the universe, to see what led us to support or reject different models of its large scale structure.

#### **Powering the Sun**

In the 1860s, Lord Kelvin, for whom the absolute temperature scale is named, did a calculation of the age of the sun. Following a suggestion by Helmholtz, Kelvin assumed that the most powerful source of energy available to the sun was its gravitational potential energy. Noting the rate at which the sun was radiating energy, Kelvin estimated that the sun was no older than half a billion years. This was a serious problem for Darwin, whose theory of evolution required considerably longer times for the processes of evolution to have taken place. During their lifetimes neither Darwin or Kelvin could explain the apparent discrepancy of having fossils older than the sun.

This problem was overcome by the discovery that the main source of energy of the sun was not gravitational potential energy, but instead the nuclear energy released by the fusion of hydrogen nuclei to form helium nuclei. In 1938 Hans Beta worked out the details of how this process worked. The reaction begins when two protons collide with sufficient energy to overcome the Coulomb repulsion and get close enough to feel the

very strong, but short range, attractive nuclear force. Such a strong collision is required to overcome the Coulomb barrier, that fusion is a rare event in the lifetime of any particular solar proton. On the average, a solar proton can bounce around about 30 million years before fusing. There are, of course, many protons in the sun, so that many such fusions are occurring at any one time.

Just after two protons fuse, electric potential energy is released when one of the protons decays via the weak interaction into a neutron, electron, and a neutrino. The electron and neutrino are ejected, leaving behind a deuterium nucleus consisting of a proton and a neutron. This reaction is the source of the neutrinos radiated by the sun.

Within a few seconds of its creation, the deuterium nucleus absorbs another proton to become a helium 3 nucleus. Since helium 3 nuclei in the sun are quite rare, it is on the average several million years before the helium 3 nucleus collides with another helium 3 nucleus. The result of this collision is the very stable helium 4 nucleus and the ejection of 2 protons. The net result of all these steps is the conversion of 4 protons into a helium 4 nucleus with the release of .6% of the proton's rest mass energy in the form of neutrinos and photons.

Not only did Beta's theory provide an explanation for the source of the sun's energy, it also demonstrated how elements can be created inside of a star. It raised the question of whether all the elements could be created inside stars. Could you start with stars initially containing only hydrogen gas and end up with all the elements we see around us?

#### Abundance of the Elements

From studies of minerals in the earth and in meteorites, and as a result of astronomical observations, we know considerable detail about the abundances of the elements around us. As seen in the chart of Figure (9), hydrogen and helium are the most abundant elements, followed by peaks at carbon, oxygen, iron and lead. There is a noticeable lack of lithium, beryllium and boron, and a general trailing off of the heavier elements. Is it possible to explain not only how elements could be created in stars, but also explain these observed abundances as being the natural result of the nuclear reactions inside stars?

The first problem is the fact that there are no stable nuclei with 5 or 8 nucleons. This means you cannot form a stable nucleus either by adding one proton to a helium 4 nucleus or fusing two helium 4 nuclei. How, then, would the next heavier element be formed in a star that consisted of only hydrogen and helium 4? The

10<sup>12</sup>

answer was supplied by E. E. Salpeter in 1952 who showed that two helium 4 nuclei could produce an unstable beryllium 8 nucleus. In a dense helium rich stellar core, the beryllium 8 nucleus could, before decaying, collide with another helium 4 nucleus forming a stable carbon 12 nucleus. One result is that elements between helium and carbon are skipped over in the element formation process, explaining the exceptionally low cosmic abundances of lithium, beryllium and boron seen in Figure (10).

The biggest barrier to explaining element formation in stars is the fact that the iron 56 nucleus is the most stable of all nuclei. Energy is released if the small nuclei fuse together to create larger ones, but energy is also released if the very largest nuclei split up (as in the case of the fission of uranium in an atomic bomb or nuclear reactor). To put it another way, energy is released making nuclei up to iron, but it costs energy to build nuclei larger than iron. Iron is the ultimate ash of nuclear reactions. How then could elements heavier than iron be created in the nuclear furnaces of stars?



**Figure 10** *Abundance of the elements* 

In 1956 the element technetium 99 was identified in the spectra of a certain class of stars. Technetium 99, heavier than iron, is an unstable element with a half life of only two hundred thousand years. On a cosmic time scale, this element had to have been made quite recently. Thus elements heavier than iron are now being created in some kind of a process.

Soon after the observation of technetium, the British astronomer Geoffrey Burbidge, looking over recently declassified data from the Bikini Atoll hydrogen bomb tests, noticed that one of the elements created in the explosion, californium 254, had a half life of 55 days. Burbidge realized that this was also the half life of the intensity of a recently observed supernova explosion. This suggested that the light from the supernova was powered by decaying californium 254. That meant that it was the supernova explosion itself that created the very heavy californium 254, and probably all the other elements heavier than iron.

In 1957 Geoffrey and Margaret Burbidge, along with the nuclear physicist William Fowler at Caltech and the British astronomer Fred Hoyle, published a famous paper showing how the fusion process in stars could explain the abundances of elements up to iron, and how supernova explosions could explain the formation of elements heavier than iron. This was one of the important steps in the use of our knowledge of the behavior of matter on a small scale, namely nuclear physics, to explain what we see on a large scale—the cosmic abundance of the elements.

# The Steady State Model of the Universe

A model of the universe, proposed in 1948 by Fred Hoyle, Herman Bondi and Thomas Gold, fit very well with the idea that all the elements in the universe heavier than hydrogen, were created as a result of nuclear reactions inside stars. This was the so-called *steady state* model. Knowing that the universe is expanding, it seems to be a contradiction to propose that the universe is steady state—i.e., that on the average, it is unchanging. If the universe is expanding and galaxies are flying apart, then in a few billion years the galaxies will be farther separated from each other than they are today. This is hardly a steady state picture.

The steady state theory got around this problem by proposing that matter was continually being created to replace that being lost due to the expansion. Consider, for example, a sphere a billion light years in diameter, centered on the earth. Over the next million years a certain number of stars will leave the sphere due to the expansion. To replace this matter flowing out of the sphere, the steady state theory assumed that hydrogen atoms were continually being created inside the sphere. All that was needed was about one hydrogen atom to be created in each cubic kilometer of space every year.

The advantage of constructing a model like the steady state theory is that the model makes certain definite predictions that can be tested. One prediction is that all the matter around us originated in the form of the hydrogen atoms that are assumed to be continually created. This implies that the heavier elements we see around us must be created by ongoing processes such as nuclear reactions inside stars. This provided a strong incentive for Hoyle and others to see if nuclear synthesis inside stars, starting from hydrogen, could explain the observed abundance of elements.

Another prediction of the steady state model is that galaxies far away must look much like nearby galaxies. When you look far away, you are also looking back in time. If you look at a galaxy one billion light years away, you are seeing light that started out a billion years ago. Light reaching us from a galaxy 10 billion light years away started out 10 billion light years ago. If the universe is really in a steady state, then galaxies 10 billion years ago should look much like galaxies do today.

# THE BIG BANG MODEL

The discovery of the expansion of the universe suggests another model of the universe, namely that the universe started in one gigantic explosion, and that the expansion we now see is the result of the pieces from that explosion flying apart.

To see why you are led to the idea of an explosion, imagine that you take a motion picture of the expanding universe and then run the motion picture backwards. If the expansion is uniform, then in the reversed motion picture we see a uniform contraction. The particles in this picture are the galaxies which are getting closer and closer together. There is a time, call it t = 0, when all the galaxies come together at a point. Now run the motion picture forward and the galaxies all move out as if there were an explosion at that point.

The explosion of the universe was first proposed by the Belgian priest and mathematician Georges Lemaître in the late 1920s. It was, in fact, Lemaître who explained Hubble's red shift versus distance data as evidence for the expansion of the universe. In the late 1920s not much was known about nuclear physics, even the neutron had not yet been discovered. But in the 1940s after the development of the atomic fission bomb and during the design of the hydrogen fusion bomb, physicists gained considerable experience with nuclear reactions in hot, dense media, and some, George Gamov in particular, began to explore the consequences of the idea that the universe started in an initial gigantic explosion.

A rough picture of the early universe in the explosion model can be constructed using the concepts of the Doppler effect and thermal equilibrium. Let us see how this works.

We have seen that the red shift of the spectral lines of light from distant galaxies can be interpreted as being caused by the stretching of the wavelengths of the light due to the expansion of the universe. In a reverse motion picture of the universe, distant galaxies would be coming toward us and the wavelengths of the spectral lines would be blue shifted. We would say that the universe was contracting, shrinking the wavelength of the spectral lines. The amount of contraction would depend upon how far back toward the t = 0 origin we went. If we went back to when the universe was 1/10

as big as it is now, wavelengths of light would contract to 1/10 their original size.

In the Einstein photoelectric effect formula,  $E_{photon} = hf = hc/\lambda$ , the shorter the photon wavelength, the more energetic the photons become. This suggests that as we compress the universe in the time reversal moving picture, photon energies increase. If there is no limit to the compression, then there is no limit to how much the photon energies increase.

Now introduce the idea of thermal equilibrium. If we go back to a very small universe, we have very energetic photons. If these photons are in thermal equilibrium with other forms of matter, as they are inside of stars, then all of the matter has enormous thermal energy, and the temperature is very high. Going back to a zero sized universe means going back to a universe that started out at an infinite temperature. Fred Hoyle thought that this picture was so ridiculous that he gave the explosion model of the universe the derisive name the "Big Bang" model. The name has stuck.

# The Helium Abundance

In the mid 1950s the cosmological theory taken seriously by most physicists was the steady state theory. In the late 1940s George Gamov had suggested that the elements had been created in the big bang when the universe was very small, dense and hot. But the work of Hoyle and Fowler was showing that the abundance of the elements could much more satisfactorily be explained in terms of nuclear synthesis inside of stars. This nuclear synthesis also explained the energy source in stars and the various stages of stellar evolution. What need was there to propose some gigantic, cataclysmic explosion?

Hoyle soon found a need. Most of the energy released in nuclear synthesis in stars results from the burning of hydrogen to form helium. By observing how much energy is released by stars, you can estimate how much helium should be produced. By the early 1960s Hoyle began to realize that nuclear synthesis could not produce enough helium to explain the observed cosmic abundance of 25%. In a 1964 paper with R. J. Taylor, Hoyle himself suggested that perhaps much of the helium was created in an initial explosion of the universe.

#### **Cosmic Radiation**

In a talk given at Johns Hopkins in early 1965, Princeton theoretician P. J. E. Peebles suggested that the early universe must have contained a considerable amount of radiation if the big bang model were correct. If there were little radiation, any hydrogen present in the early universe would have quickly fused to form heavier elements, and no hydrogen would be left today. This directly contradicts the observation that about 75% of the matter we see today consists of hydrogen. If, however, there were a large amount of radiation present in the early universe, the energetic photons would bust up the larger nuclei as they formed, leaving behind hydrogen.

Peebles proposed that this radiation, the cosmic photons which prevented the fusion of hydrogen in the early universe still exist today but in a very altered form. There should have been little change in the number of photons, but a great change in their energy. As the universe expanded, the wavelength of the cosmic photons should be stretched by the expansion, greatly reducing their energy. If the photons were in thermal equilibrium with very hot matter in the early universe, they should still have a thermal black body spectrum, but at a much lower temperature. He predicted that the temperature of the cosmic radiation should have dropped to around 10 kelvin. His colleagues at Princeton, P. G. Roll and D.T. Wilkinson were constructing a special antenna to detect such radiation. All of this work had been suggested by R. H. Dicke, inventor of the key microwave techniques needed to detect ten degree photons.

Peebles was not the first to suggest that there should be radiation left over from the big bang. That was first suggested in a 1948 paper by George Gamov and colleagues Ralph Alpher and Robert Herman in a model where all elements were to be created in the big bang. A more realistic model of the big bang proposed by Alpher and Herman in 1953 also led to the same prediction of cosmic radiation. In both cases, it was estimated that the thermal radiation should now have a temperature of 5 kelvin. In the early 1950s, Gamov, Alpher and Herman were told by radio astronomers that such radiation could not be detected by equipment then available, and the effort to detect it was not pursued. Peebles was unaware of these earlier predictions.

# THE THREE DEGREE RADIATION

In 1964, two radio astronomers working for Bell Labs, Arno Penzias and Robert Wilson, began a study of the radio waves emitted from parts of our galaxy that are away from the galactic plane. They expected a faint diffuse radiation from this part of the galaxy and planned to use a sensitive low noise radio antenna shown in Figure (11), an antenna left over from the *Echo* satellite experiment. (In that early experiment on satellite communication, a reflecting balloon was placed in orbit. The low noise antenna was built to detect the faint radio signals that bounced off the balloon.)

Since the kind of signals Penzias and Wilson expected to detect would look a lot like radio noise, they had to be careful that the signals they recorded were coming from the galaxy rather than from noise generated by the antenna or by electronics. To test the system, they looked for signals at a wavelength of 7.35 cm, a wavelength where the galaxy was not expected to produce much radiation. They found, however, a stronger signal than expected. After removing a pair of pigeons that were living in the antenna throat, cleaning out the nest and other debris which Wilson referred to as "a white dielectric material", and taking other steps to eliminate noise, the extra signal persisted.



**Figure 11** *Penzias and Wilson, and the Holmdel radio telescope.* 

If the 7.35 cm wavelength signal were coming from the galaxy, there should be regions of the galaxy that produced a stronger signal than other regions. And the neighboring galaxy Andromeda should also be a localized source of this signal. However Penzias and Wilson found that the 7.35 cm signal was coming in uniformly from all directions. The radiation had to be coming in from a much larger region of space than our galaxy.

Studies of the signal at still shorter wavelengths showed that if the signal were produced by a blackbody spectrum of radiation, the effective temperature would be about 3.5 kelvin. Penzias talked with a colleague who had talked with another colleague who had attended Peebles' talk at Johns Hopkins on the possibility of radiation left over from the big bang. Penzias and Wilson immediately suspected that the signal they were detecting might be from this radiation.

Penzias and Wilson could detect only the long wavelength tail and of the three degree radiation. Three degree radiation should have a maximum intensity at a wavelength given by the Wein formula, Equation 1,

$$\lambda_{\max} = \frac{2.898 \text{ mm K}}{\text{T}}$$

$$= \frac{2.898 \text{ mm K}}{3\text{K}} \approx 1 \text{ mm}$$
(17)

Radiation with wavelengths in the 1 mm region cannot get through the earth's atmosphere. As a result Penzias and Wilson, and others using ground based antennas, could not verify that the radiation had a complete blackbody spectrum. From 1965 to the late 1980s, various balloon and rocket based experiments, which lifted antennas above the earth's atmosphere, verified that the radiation detected by Penzias and Wilson was part of a complete blackbody spectrum of radiation at a temperature of 2.74 kelvin.

In 1989, NASA orbited the *COBE* (Cosmic Background Explorer) satellite to make a detailed study of the cosmic background radiation. The results from this satellite verified that this radiation has the most perfect blackbody spectrum ever seen by mankind. The temperature is 2.735 kelvin with variations of the order of one part in 100,000. The questions we have to deal with now are not whether there is light left over from the big bang, but why it is such a nearly perfect blackbody spectrum.

# **Thermal Equilibrium of the Universe**

That the cosmic background radiation has nearly a perfect blackbody spectrum tells us that at some point in its history, the universe was in nearly perfect thermal equilibrium, with everything at one uniform temperature. That is certainly not the case today. The cosmic radiation is at a temperature of 2.735 kelvin, Hawaii has an average temperature of 295 kelvin, and the temperature inside of stars ranges up to billions of degrees. There must have been a dramatic change in the nature of the universe sometime in the past. That change occurred when the universe suddenly became transparent at an age of about 700,000 years.

To see why the universe suddenly became transparent, and why this was such an important event, it is instructive to reconstruct what the universe must have been like at still earlier times.

# THE EARLY UNIVERSE

Imagine that we have a videotape recording of the evolution of the universe. We put the tape in our VCR and see that the tape has not been rewound. It is showing our current universe with stars, galaxies and the cosmic radiation at a temperature of 2.735 k. You can calculate the density of photons in the cosmic radiation, and compare that with the average density of protons and neutrons (nucleons) in the stars and galaxies. You find that the photons outnumber the nucleons by a factor of about 10 billion to 1. Although there are many more photons than nucleons, the rest energy of a proton or neutron is so much greater than the energy of a three degree photon that the total rest energy of the stars and galaxies is about 100 times greater than the total energy in the cosmic radiation.

Leaving the VCR on play, we press the rewind button. The picture is not too clear, but we can see general features of the contracting universe. The galaxies are moving together and the wavelength of the cosmic radiation is shrinking. Since the energy of the cosmic photons is given by Einstein's formula  $E = hc/\lambda$ , the shrinking of the photon wavelengths increases their energy. On the other hand the rest mass energy of the stars and galaxies is essentially unaffected by the contraction of the universe. As a result the energy of the cosmic photons is becoming a greater and greater share of the total energy of the universe. When the universe has contracted to about 1/100th of its present size, when the universe is about 1/2 million years old, the cosmic photons have caught up to the matter particles. At earlier times, the cosmic photons have more energy than other forms of matter.

# **The Early Universe**

As the tape rewinds our attention is diverted. When we look again at the screen, we see that the tape is showing a very early universe. The time indicated is .01 seconds! The temperature has risen to 100 billion degrees, and the thermal photons have an average energy of 40 million electron volts! We obviously missed a lot in the rewind. Stopping the tape, we then run it forward to see what the universe looks like at this very early stage. There is essentially the same number of nucleons in this early universe as there are today. Since the thermal energy of 40 MeV is much greater than the 1.3 MeV mass difference between neutrons and protons, there is enough thermal energy to freely convert protons into neutrons, and vice versa. As a result there are about equal numbers of protons and neutrons. There is also about the same number of thermal photons in this early universe as there are today, about 100 billion photons for each nucleon.

While there is not much change from today in the number of nucleons or photons in our .01 second universe, there is a vastly different number of electrons. The thermal photons, with an average energy of 40 MeV, can freely create positron and electron pairs. The rest energy of a positron or an electron is only .5 MeV, thus only 1 MeV is required to create a pair. The result is that the universe at this time is a thermal soup of photons, positrons and electrons—about equal numbers. There are also many neutrinos left over from an earlier time. All of those species outnumber the few nucleons by a factor of about 100 billion to one.

#### **Excess of Matter over Antimatter**

If you look closely and patiently count the number of positrons and electrons in some region of space, you will find that for every 100,000,000,000 positrons, there are 100,000,000,001 electrons. The electrons outnumber the positrons by 1 in 100 billion. In fact, the excess number of negative electrons is just equal to the number of positive protons, with the result that the universe is electrically neutral.

The tiny excess of electrons over positrons represents an excess of matter over antimatter. In most particle reactions we study today, if particles are created, they are created in particle, antiparticle pairs. The question is then, why does this early universe have a tiny excess of matter particles over antimatter particles? What in the still earlier universe created this tiny imbalance? There is a particle reaction, caused by the weak interaction, that does not treat matter and antimatter symmetrically. This reaction, discovered by Val Fitch in 1964, could possibly explain how this tiny imbalance came about. It is not clear whether there was enough time in the very early universe for Fitch's reaction to create the observed imbalance. An excellent guidebook for our video tape is Steven Weinberg's *The First Three Minutes*. Weinberg was one of the physicists who discovered the connection between the weak interaction and electromagnetism. Weinberg breaks up the first three minutes of the life of the universe into five frames. We happened to have stopped the tape recording at Weinberg's frame #1. To see what we missed in our fast rewind, we will now run the tape forward, picking up the other four frames in the first three minutes as well as important later events.

# Frame #2 (.11 seconds)

As we run the tape forward, the universe is now expanding, the wavelength of the thermal photons is getting longer, and their temperature is dropping. When the time counter gets up to t = .11 seconds, the temperature has dropped to 30 billion kelvin and the average energy of the thermal photons has dropped to 10 MeV. Back at frame #1, when the thermal energy was 40 MeV, there were roughly equal numbers of protons and neutrons. However, the lower thermal energy of 10 MeV is not sufficiently greater than the 1.3 MeV proton-neutron mass difference to maintain the equality.

In the many rapid collisions where protons are being converted into neutrons and vice versa (via the weak interaction), there is a slightly greater chance that the heavier neutron will decay into a lighter proton rather than the other way around. As a result the percentage of neutrons has dropped to 38% by the time t = .11 seconds.

# Frame #3 (1.09 seconds)

Aside from the drop in temperature and slight decrease in the percentage of neutrons, not much else happened as we went from frame #1 at .01 seconds to frame #2 at .11 seconds. Starting up the tape player again, we go forward to t = 1.09 seconds, Weinberg's third frame. The temperature has dropped to 10 billion kelvin, which corresponds to a thermal energy of 4 MeV. This is not too far above the 1 MeV threshold for creating positron electron pairs. As a result the positron electron pairs are beginning to annihilate faster than they are being created. Also by this time the percentage of neutrons has dropped to 24%.

#### Frame #4 (13.82 seconds)

At a time of 13.82 seconds, Weinberg's fourth frame, the temperature has dropped to 3 billion kelvin, corresponding to an average thermal energy of 1 MeV per particle. With any further drop in temperature, the average thermal photon will not have enough energy to create positron electron pairs. The result is that vast numbers of positrons and electrons are beginning to annihilate each other. Soon there will be equal numbers of electrons and protons, and the only particles remaining in very large numbers will be neutrinos and thermal photons.

By this fourth frame, the percentage of neutrons has dropped to 17%. The temperature of 3 billion degrees is low enough for helium nuclei to survive, but helium nuclei do not form because of the deuterium bottleneck. When a proton and neutron collide, they can easily form a deuterium nucleus. Although deuterium is stable, it is weakly bound. At a temperature of 3 billion kelvin, the thermal protons quickly break up any deuterium that forms. Without deuterium, it is not possible to build up still larger nuclei.

# Frame #5 (3 minutes and 2 seconds)

Going forward to a time of 3 minutes and 2 seconds, the universe has cooled to a billion kelvin, the positrons and most electrons have disappeared, and the only abundant particles are photons, neutrinos and antineutrinos. The neutron proton balance has dropped to 14% neutrons. While tritium (one proton and two neutrons) and helium 4 are stable at this temperature, deuterium is not, thus no heavier nuclei can form.

A short time later, the temperature drops to the point where deuterium is stable. When this happens, neutrons can combine with protons to form deuterium and tritium, and these then combine to form helium 4. Almost immediately the remaining nearly 13% neutrons combine with an equal number of protons to form most of the 25% abundance of cosmic helium we see today. This is where the helium came from that Hoyle could not explain in terms of nuclear synthesis inside of stars.

34-31

Because there are no stable nuclei with 5 or 8 nucleons, there is no simple route to the formation of still heavier elements. At a temperature of a billion degrees, the universe is only about 70 times hotter than the core of our sun, cooler than the core of hot stars around today that are fusing the heavier elements. As a result, nuclear synthesis in the early universe stops at helium 4 with a trace of lithium 7. One of the best tests of the big bang theory is a rather precise prediction of the relative abundances of hydrogen, deuterium, helium 4 and lithium 7, all left over from the early universe. When the formation of these elements is complete, the universe is 3 minutes and 46 seconds old.

# Decoupling (700,000 years)

Continue running the tape forward, and nothing of much interest happens for a long time. The thermal photons still outnumber the nucleons and electrons by a factor of about 10 billion to one, and the constant collisions between these particles prevent the formation of atoms. What we see is a hot, ionized, nearly uniform plasma consisting of photons, charged nuclei and separate electrons. As time goes on, the plasma is expanding and cooling.

When you look at the sun, you see a round ball with an apparently sharp edge. But the sun is not a solid object with a well defined surface. Instead, it is a bag of mostly hydrogen gas held together by gravity. It is hottest at the center and cools off as you go out from the center. At what appears to us to be the surface, the temperature has dropped to about 3,000 kelvin.

At a temperature above 3,000 kelvin, hydrogen gas becomes ionized, a state where an appreciable fraction of the electrons are torn free from the proton nuclei. When the gas is ionized, it is opaque because photons can interact directly with the free charges present in the gas. Below a temperature of 3,000 kelvin, hydrogen consists essentially of neutral atoms which are unaffected by visible light. As a result the cooler hydrogen gas is transparent. The apparent surface of the sun marks the abrupt transition from an opaque plasma, at temperatures above 3,000K, to a transparent gas at temperatures below 3,000K. A similar transition takes place in the early universe. By the age of about 700,000 years, the universe cools to a temperature of 3,000K. Before that the universe is an opaque plasma like the inside of the sun. The photons in thermal equilibrium with the matter particles have enough energy to bust up any complete atoms and any gravitational clumps that are trying to form.

When the universe drops to a temperature below 3,000 kelvin, the hydrogen gas forms neutral atoms and becomes transparent. (The 25% helium had already become neutral some time earlier). As a result the universe suddenly becomes transparent, and the thermal photons decouple from matter.

From this decoupling on, there is essentially no interaction between the thermal photons and any form of matter. All that happens to the photons is that their wavelength is stretched by the expansion of the universe. This stretching preserves the blackbody spectrum of the photons while lowering the effective blackbody temperature. This blackbody spectrum is now at the temperature of 2.735K, as observed by the COBE satellite.

When the matter particles are decoupled, freed from the constant bombardment of the cosmic photons, gravity can begin the work of clumping up matter to form stars, globular clusters, black holes and galaxies. All these structures start to form after the decoupling, after the universe is 700,000 years old. It is this formation of stars and galaxies that we see as we run the tape forward to our present day.

Looking out with ever more powerful telescopes is essentially equivalent to running our videotape backwards. The farther out we look, the farther back in time we see. Images from the Hubble telescope are giving us a view back toward the early universe when galaxies were very much younger and quite different than they are today. The most distant galaxy we have identified so far emitted light when the universe was 5% of its current size. What happens when we build still more powerful telescopes and look still farther back? When we look out so far that the universe is only 700,000 years old, we are looking at the universe that has just become transparent. *We can see no farther*! To look farther is like trying to look down inside the surface of the sun.

In fact we do not need a more powerful telescope to see this far back. The three degree cosmic background radiation gives us a fantastically clear, detailed photograph of the universe at the instant it went transparent.

The horn antenna used by Penzias and Wilson was the first device to look at a small piece of this photograph. The COBE satellite looked at the whole photograph, but with rather limited resolution. COBE detected some very tiny lumpiness, temperature variations of about one part in 100,000. This lumpiness may have been what gravity needed to start forming galaxies. A higher resolution photograph will be needed to tell for sure.

# Guidebooks

We ran the videotape quite rapidly without looking at many details. Our focus has been on the formation of the elements and the three degree radiation, two of the main pieces of evidence for the existence of a big bang. We have omitted a number of fascinating details such as how dense was the early universe, when did the neutrinos decouple from matter, and what happened before the first frame? There are excellent guidebooks that accompany this tape where you can find these details. There is Weinberg's The First Three Minutes which we have mentioned. The 1993 edition has an addendum that introduces some ideas about the very, very early universe, when the universe was millions of times younger and hotter than the first frame. Perhaps the best guidebook to how mankind came to our current picture of the universe is the book by Timothy Ferris Coming of Age in the Milky Way. Despite the title, this is one of the most fascinating and readable accounts available. In our discussion we have drawn much from Weinberg and Ferris.

# Chapter 35 Bohr Theory of Hydrogen

The hydrogen atom played a special role in the history of physics by providing the key that unlocked the new mechanics that replaced Newtonian mechanics. It started with Johann Balmer's discovery in 1884 of a mathematical formula for the wavelengths of some of the spectral lines emitted by hydrogen. The simplicity of the formula suggested that some understandable mechanisms were producing these lines.

The next step was Rutherford's discovery of the atomic nucleus in 1912. After that, one knew the basic structure of atoms—a positive nucleus surrounded by negative electrons. Within a year Neils Bohr had a model of the hydrogen atom that "explained" the spectral lines. Bohr introduced a new concept, the energy level. The electron in hydrogen had certain allowed energy levels, and the sharp spectral lines were emitted when the electron jumped from one energy level to another. To explain the energy levels, Bohr developed a model in which the electron had certain allowed orbits and the jump between energy levels corresponded to the electron moving from one allowed orbit to another.

Bohr's allowed orbits followed from Newtonian mechanics and the Coulomb force law, with one small but crucial modification of Newtonian mechanics. The angular momentum of the electron could not vary continuously, it had to have special values, be quantized in units of Planck's constant divided by  $2\pi$ , h/ $2\pi$ . In Bohr's theory, the different allowed orbits corresponded to orbits with different allowed values of angular momentum.

Again we see Planck's constant appearing at just the point where Newtonian mechanics is breaking down. There is no way one can explain from Newtonian mechanics why the electrons in the hydrogen atom could have only specific quantized values of angular momentum. While Bohr's model of hydrogen represented only a slight modification of Newtonian mechanics, it represented a major philosophical shift. Newtonian mechanics could no longer be considered the basic theory governing the behavior of particles and matter. Something had to replace Newtonian mechanics, but from the time of Bohr's theory in 1913 until 1924, no one knew what the new theory would be.

In 1924, a French graduate student, Louis de Broglie, made a crucial suggestion that was the key that led to the new mechanics. This suggestion was quickly followed up by Schrödinger and Heisenberg who developed the new mechanics called **quantum mechanics**. In this chapter our focus will be on the developments leading to de Broglie's idea.

# THE CLASSICAL HYDROGEN ATOM

With Rutherford's discovery of the atomic nucleus, it became clear that atoms consisted of a positively charged nucleus surrounded by negatively charged electrons that were held to the nucleus by an electric force. The simplest atom would be hydrogen consisting of one proton and one electron held together by a Coulomb force of magnitude

$$\left|\vec{F}_{e}\right| = \frac{e^{2}}{r^{2}} \qquad \stackrel{p \ \vec{F}_{e}}{\longleftarrow} r \xrightarrow{\vec{F}_{e}} e \qquad (1)$$

(For simplicity we will use CGS units in describing the hydrogen atom. We do not need the engineering units, and we avoid the complicating factor of  $1/4\pi\epsilon_0$  in the electric force formula.) As shown in Equation 1, both the proton and the electron attract each other, but since the proton is 1836 times more massive than the electron, the proton should sit nearly at rest while the electron orbits around it.

Thus the hydrogen atom is such a simple system, with known masses and known forces, that it should be a straightforward matter to make detailed predictions about the nature of the atom. We could use the orbit program of Chapter 8, replacing the gravitational force  $GMm/r^2$  by  $e^2/r^2$ . We would predict that the electron moved in an elliptical orbit about the proton, obeying all of Kepler's laws for orbital motion.

There is one important point we would have to take into account in our analysis of the hydrogen atom that we did not have to worry about in our study of satellite motion. The electron is a charged particle, and accelerated charged particles radiate electromagnetic waves. Suppose, for example, that the electron were in a circular orbit moving at an angular velocity  $\omega$  as shown in Figure (1a). If we were looking at the orbit from the side, as shown in Figure (1b), we would see an electron oscillating up and down with a velocity given by  $v = v_0 \sin \omega t$ .

In our discussion of radio antennas in Chapter 32, we saw that radio waves could be produced by moving electrons up and down in an antenna wire. If electrons oscillated up and down at a frequency  $\omega$ , they produced radio waves of the same frequency. Thus it is a prediction of Maxwell's equations that the electron in the hydrogen atom should emit electromagnetic radiation, and the frequency of the radiation should be the frequency at which the electron orbits the proton.

For an electron in a circular orbit, predicting the motion is quite easy. If an electron is in an orbit of radius r, moving at a speed v, then its acceleration  $\vec{a}$  is directed toward the center of the circle and has a magnitude

$$\vec{a} = \frac{v^2}{r} \tag{2}$$

Using Equation 1 for the electric force and Equation 2 for the acceleration, and noting that the force is in the same direction as the acceleration, as indicated in Figure (2), Newton's second law gives

$$\begin{vmatrix} \vec{F} \\ = m \begin{vmatrix} \vec{a} \end{vmatrix}$$
$$\frac{e^2}{r^2} = m \frac{v^2}{r}$$
(3)

One factor of r cancels and we can immediately solve for the electron's speed v to get  $v^2 = e^2/mr$ , or

$$v_{\text{electron}} = \frac{e}{\sqrt{\mathrm{mr}}}$$
 (4)

The period of the electron's orbit should be the distance  $2\pi r$  travelled, divided by the speed v, or  $2\pi r/v$  seconds per cycle, and the frequency should be the inverse of that, or  $v/2\pi r$  cycles per second. Using Equation 4 for v, we get

frequency of electron in orbit = 
$$\frac{v}{2\pi r} = \frac{e}{2\pi r\sqrt{mr}}$$
 (5)

According to Maxwell's theory, this should also be the frequency of the radiation emitted by the electron.



Figure 1

The side view of circular motion is an up and down oscillation.

Electromagnetic radiation carries energy. Thus, to see what effect this has on the electron's orbit, let us look at the formula for the energy of an orbiting electron.

From Equation 3 we can immediately solve for the electron's kinetic energy. The result is

$$\frac{1}{2}mv^2 = \frac{e^2}{2r} \quad \stackrel{electron}{\underset{energy}{kinetic}} \tag{6}$$

The electron also has electric potential energy just as an earth satellite had gravitational potential energy. The formula for the gravitational potential energy of a satellite was

potential energy  
of an earth satellite = 
$$-\frac{GMm}{r}$$
 (10-50a)

where M and m are the masses of the earth and the satellite respectively. This is the result we used in Chapter 8 to test for conservation of energy (Equations 8-29 and 8-31) and in Chapter 10 where we calculated the potential energy (Equations 10-50a and 10-51). The minus sign indicated that the gravitational force is attractive, that the satellite starts with zero potential energy when  $r = \infty$  and loses potential energy as it falls in toward the earth.

We can convert the formula for gravitational potential energy to a formula for electrical potential energy by comparing formulas for the gravitational and electric forces on the two orbiting objects. The forces are

$$\vec{F}_{\text{gravity}} = \frac{\text{GMm}}{r^2}$$
;  $|\vec{F}_{\text{electric}}| = \frac{e^2}{r^2}$ 

Since both are  $1/r^2$  forces, we can go from the gravitational to the electric force formula by replacing the



Figure 2 For a circular orbit, both the acceleration a and the force F point toward the center of the circle. Thus we can equate the magnitudes of F and ma.

constant GMm by  $e^2$ . Making this same substitution in the potential energy formula gives

$$PE = \frac{-e^2}{r} \qquad \begin{array}{c} electrical potential energy\\ of the electron in the\\ hydrogen atom \end{array}$$
(7)

Again the potential energy is zero when the particles are infinitely far apart, and the electron loses potential energy as it falls toward the proton. (We used this result in the analysis of the binding energy of the hydrogen molecule ion, explicitly in Equation 18-15.)

The formula for the total energy  $E_{total}$  of the electron in hydrogen should be the sum of the kinetic energy, Equation 6, and the potential energy, Equation 7.

$$E_{total} = \underset{energy}{\text{kinetic}} + \underset{energy}{\text{potential}}$$
$$= \frac{e^2}{2r} - \frac{e^2}{r}$$
$$E_{total} = \frac{-e^2}{2r} \qquad \underset{of \ electron}{total \ energy} \tag{8}$$

The significance of the minus (–) sign is that the electron is bound. Energy is required to pull the electron out, to ionize the atom. For an electron to escape, its total energy must be brought up to zero.

We are now ready to look at the predictions that follow from Equations 5 and 8. As the electron radiates light it must lose energy and its total energy must become more negative. From Equation 8 we see that for the electron's energy to become more negative, the radius r must become smaller. Then Equation 5 tells us that as the radius becomes smaller, the frequency of the radiation increases. We are lead to the picture of the electron spiraling in toward the proton, radiating even higher frequency light. There is nothing to stop the process until the electron crashes into the proton. It is an unambiguous prediction of Newtonian mechanics and Maxwell's equations that the hydrogen atom is unstable. It should emit a continuously increasing frequency of light until it collapses.

#### **Energy Levels**

By 1913, when Neils Bohr was trying to understand the behavior of the electron in hydrogen, it was no surprise that Maxwell's equations did not work at an atomic scale. To explain blackbody radiation and the photoelectric effect, Planck and Einstein were led to the picture that light consists of photons rather than Maxwell's waves of electric and magnetic force.

To construct a theory of hydrogen, Bohr knew the following fact. Hydrogen gas at room temperature emits no light. To get radiation, it has to be heated to rather high temperatures. Then you get distinct spectral lines rather than the continuous radiation spectrum expected classically. The visible spectral lines are the  $H_{\alpha}$ ,  $H_{\beta}$  and  $H_{\gamma}$  lines we saw in the hydrogen spectrum experiment. These and many infra red lines we saw in the spectrum of the hydrogen star, Figure (33-28) reproduced below, make up the Balmer series of lines. Something must be going on inside the hydrogen atom to produce these sharp spectral lines.

Viewing the light radiated by hydrogen in terms of Einstein's photon picture, we see that the hydrogen atom emits photons with certain precise energies. As an exercise in the last chapter you were asked to calculate, in eV, the energies of the photons in the  $H_{\alpha}$ ,  $H_{\beta}$  and  $H_{\gamma}$  spectral lines. The answers are

$$E_{H_{\alpha}} = 1.89 \text{ eV}$$
  
 $E_{H_{\beta}} = 2.55 \text{ eV}$   
 $E_{H_{\gamma}} = 2.86 \text{ eV}$  (9)

 $3.75 \times 10^{-5}$ 

H12

The question is, why does the electron in hydrogen emit only certain energy photons? The answer is Bohr's main contribution to physics. Bohr assumed that the electron had, for some reason, only certain allowed energies in the hydrogen atom. He called these *allowed energy levels*. When an electron jumped from one energy level to another, it emitted a photon whose energy was equal to the difference in the energy of the two levels. The red 1.89 eV photon, for example, was radiated when the electron fell from one energy level to another level 1.89 eV lower. There was a bottom, lowest energy level below which the electron could not fall. In cold hydrogen, all the electrons were in the bottom energy level and therefore emitted no light.

When the hydrogen atom is viewed in terms of Bohr's energy levels, the whole picture becomes extremely simple. The lowest energy level is at -13.6 eV. This is the total energy of the electron in any cold hydrogen atom. It requires 13.6 eV to ionize hydrogen to rip an electron out.

Figure 3 Energy level diagram for the hydrogen atom. All the energy levels are given by the simple formula  $E_n = -13.6/n^2 eV.$ All Balmer series lines result from jumps down to the n = 2 level. The 3 jumps shown give rise to the three visible hydrogen lines.

 $3.85 \times 10^{-5}$ 

H9

 $3.80 \times 10^{-5}$ 

| H10

wavelength

Н11



**Figure 33-28** Spectrum of a hydrogen star

| H20  $3.70 \times 10^{-5}$ 

H15 H14

| H13

3.65 × 10<sup>-5</sup>

| | H40 H30 The first energy level above the bottom is at -3.40 eV which turns out to be (-13.6/4) eV. The next level is at -1.51 eV which is (-13.6/9) eV. All of the energy levels needed to explain every spectral line emitted by hydrogen are given by the formula

$$E_n = \frac{-13.6 \text{ eV}}{n^2}$$
(10)

where n takes on the integer values 1, 2, 3, .... These energy levels are shown in Figure (3).

#### **Exercise 1**

Use Equation 10 to calculate the lowest 5 energy levels and compare your answer with Figure 3.

Let us see explicitly how Bohr's energy level diagram explains the spectrum of light emitted by hydrogen. If, for example, an electron fell from the n=3 to the n=2 level, the amount of energy  $E_{3-2}$  it would lose and therefore the energy it would radiate would be

$$E_{3-2} = E_3 - E_2$$
  
= -1.51 eV - (-3.40 eV)  
= 1.89 eV  
= energy lost in falling  
from n = 3 to n = 2 level (11)

which is the energy of the red photons in the  $H_{\alpha}$  line.

#### Exercise 2

Show that the H<sub> $\beta$ </sub> and H<sub> $\gamma$ </sub> lines correspond to jumps to the n = 2 level from the n = 4 and the n = 5 levels respectively.

From Exercise 2 we see that the first three lines in the Balmer series result from the electron falling from the third, fourth and fifth levels down to the second level, as indicated by the arrows in Figure (3).

All of the lines in the Balmer result from jumps down to the second energy level. For historical interest, let us see how Balmer's formula for the wavelengths in this series follows from Bohr's formula for the energy levels. For Balmer's formula, the lines we have been calling  $H_{\alpha}$ ,  $H_{\beta}$  and  $H_{\gamma}$  are  $H_3$ ,  $H_4$ ,  $H_5$ . An arbitrary line in the series is denoted by  $H_n$ , where n takes on the values starting from 3 on up. The Balmer formula for the wavelength of the  $H_n$  line is from Equation 33-6

$$\lambda_{\rm n} = 3.65 \times 10^{-5} {\rm cm} \times \left(\frac{{\rm n}^2}{{\rm n}^2 - 4}\right)$$
 (33-6)

Referring to Bohr's energy level diagram in Figure (3), consider a drop from the nth energy level to the second. The energy lost by the electron is  $(E_n - E_2)$  which has the value

$$E_n - E_2 = \frac{13.6 \text{ eV}}{n^2} - \frac{13.6 \text{ eV}}{2^2} \qquad \begin{array}{c} \text{energy lost by} \\ \text{electron going} \\ \text{from nth to} \\ \text{second level} \end{array}$$

This must be the energy  $E(H_n)$  carried out by the photon in the  $H_n$  spectral line. Thus

$$E(H_{n}) = 13.6 \text{ eV}\left(\frac{1}{4} - \frac{1}{n^{2}}\right)$$
$$= 13.6 \text{ eV}\left(\frac{n^{2} - 4}{4n^{2}}\right)$$
(12)

We now use the formula

$$\lambda = \frac{12.4 \times 10^{-5} \text{cm} \cdot \text{eV}}{\text{E}_{\text{photon}}(\text{in eV})}$$
(34-8)

relating the photon's energy to its wavelength. Using Equation 12 for the photon energy gives

$$\lambda_{n} = \frac{12.4 \times 10^{-5} \text{cm} \cdot \text{eV}}{13.6 \text{ eV}} \left(\frac{4n^{2}}{n^{2} - 4}\right)$$
$$\lambda_{n} = 3.65 \times 10^{-5} \text{cm} \left(\frac{n^{2}}{n^{2} - 4}\right)$$

which is Balmer's formula.

It does not take great intuition to suspect that there are other series of spectral lines beyond the Balmer series. The photons emitted when the electron falls down to the lowest level, down to -13.6 eV as indicated in Figure (4), form what is called the *Lyman series*. In this series the least energy photon, resulting from a fall from -3.40 eV down to -13.6 eV, has an energy of 10.2 eV, well out in the ultraviolet part of the spectrum. All the other photons in the Lyman series have more energy, and therefore are farther out in the ultraviolet.

It is interesting to note that when you heat hydrogen and see a Balmer series photon like  $H_{\alpha}$ ,  $H_{\beta}$  or  $H_{\gamma}$ , eventually a 10.2 eV Lyman series photon must be emitted before the hydrogen can get back down to its ground state. With telescopes on earth we see many hydrogen stars radiating Balmer series lines. We do not see the Lyman series lines because these ultraviolet photons do not make it down through the earth's atmosphere. But the Lyman series lines are all visible using orbiting telescopes like the Ultraviolet Explorer and the Hubble telescope.

Figure 4

The Lyman series

far down as the

*level is called the "ground state".*)

consists of all jumps

down to the -13.6eV

level. (Since this is as

electron can go, this



Another series, all of whose lines lie in the infra red, is the Paschen series, representing jumps down to the n=3 energy level at -1.55 eV, as indicated in Figure (5). There are other infra red series, representing jumps down to the n=4 level, n=5 level, etc. There are many series, each containing many spectral lines. And all these lines are explained by Bohr's conjecture that the hydrogen atom has certain allowed energy levels, all given by the simple formula  $E_n = (-13.6/n^2) \text{ eV}$ . This one simple formula explains a huge amount of experimental data on the spectrum of hydrogen.

#### Exercise 3

Calculate the energies (in eV) and wavelengths of the 5 longest wavelength lines in

- (a) the Lyman series
- (b) the Paschen series

On a Bohr energy level diagram show the electron jumps corresponding to each line.

#### Exercise 4

In Figure (33-28), repeated 2 pages back, we showed the spectrum of light emitted by a hydrogen star. The lines get closer and closer together as we get to  $H_{40}$  and just beyond. Explain why the lines get closer together and calculate the limiting wavelength.

Figure 5 The Paschen series consists of all jumps down to the n = 3level. These are all in the infra red.


Where do Bohr's energy levels come from? Certainly not from Newtonian mechanics. There is no excuse in Newtonian mechanics for a set of allowed energy levels. But did Newtonian mechanics have to be rejected altogether? Planck was able to explain the blackbody radiation formula by patching up classical physics, by assuming that, for some reason, light was emitted and absorbed in quanta whose energy was proportional to the light's frequency. The reason why Planck's trick worked was understood later, with Einstein's proposal that light actually consisted of particles whose energy was proportional to frequency. Blackbody radiation had to be emitted and absorbed in quanta because light itself was made up of these quanta.

By 1913 it had become respectable, frustrating perhaps, but respectable to modify classical physics in order to explain atomic phenomena. The hope was that a deeper theory would come along and naturally explain the modifications.

What kind of a theory do we construct to explain the allowed energy levels in hydrogen? In the classical picture we have a miniature solar system with the proton at the center and the electron in orbit. This can be simplified by restricting the discussion to circular orbits. From our earlier work with the classical model of hydrogen, we saw that an electron in an orbit of radius r had a total energy E(r) given by

$$E(r) = \frac{-e^2}{2r} \qquad \begin{array}{c} \text{total energy of} \\ \text{an electron in} \\ \text{a circular orbit} \\ \text{of radius r} \end{array} \qquad (8 \text{ repeated})$$

If the electron can have only certain allowed energies  $E_n = -13.6/n^2 \text{ eV}$ , then if Equation (8) holds, the electron orbits can have only certain allowed orbits of radius  $r_n$  given by

$$E_n = \frac{-e^2}{2r_n} \tag{13}$$

The  $r_n$  are the radii of the famous Bohr orbits. This leads to the rather peculiar picture that the electron can exist in only certain allowed orbits, and when the electron jumps from one allowed orbit to another, it emits a photon whose energy is equal to the difference in energy between the two orbits. This model is indicated schematically in Figure (6).

## **Exercise 5**

From Equation 13 and the fact that  $E_1 = -13.6 \text{ eV}$ , calculate the radius of the first Bohr orbit  $r_1$ . [Hint: first convert eV to ergs.] This is known as the *Bohr radius* and is in fact a good measure of the actual radius of a cold hydrogen atom. [The answer is  $r_1 = .529 \times 10^{-8} \text{ cm} = .529 \text{\AA}$ .] Then show that  $r_n = n^2 r_1$ .

#### Figure 6

The Bohr orbits are determined by equating the allowed energy  $E_n = -13.6/n^2$  to the energy  $E_n = -e^2/2r_n$ for an electron in an orbit of radius  $r_n$ . The Lyman series represents all jumps down to the smallest orbit, the Balmer series to the second orbit, the Paschen series to the third orbit, etc. (The radii in this diagram are not to scale, the radii  $r_n$ increase in size as  $n^2$ , as you can easily show by equating the two values for  $E_n$ .)



## Angular Momentum in the Bohr Model

Nothing in Newtonian mechanics gives the slightest hint as to why the electron in hydrogen should have only certain allowed orbits. In the classical picture there is nothing special about these particular radii.

But ever since the time of Max Planck, there was a special unit of angular momentum, the amount given by Planck's constant h. Since Planck's constant keeps appearing whenever Newtonian mechanics fails, and since Planck's constant has the dimensions of angular momentum, perhaps there was something special about the electron's angular momentum when it was in one of the allowed orbits.

We can check this idea by re expressing the electron's total energy not in terms of the orbital radius r, but in terms of its angular momentum L.

We first need the formula for the electron's angular momentum when in a circular orbit of radius r. Back in Equation 4, we found that the speed v of the electron was given by

$$v = \frac{e}{\sqrt{mr}}$$
 (4 repeated)

Multiplying this through by m gives us the electron's linear momentum mv

$$mv = \frac{me}{\sqrt{mr}} = e\sqrt{\frac{m}{r}}$$
(14)

The electron's angular momentum about the center of the circle is its linear momentum mv times the lever arm r, as indicated in the sketch of Figure (7). The result is

$$L = (mv)r = \left(e\sqrt{\frac{m}{r}}\right)r$$
  
=  $e\sqrt{mr}$  (15)

where we used Equation 14 for mv.

The next step is to express r in terms of the angular momentum L. Squaring Equation 13 gives

$$L^2 = e^2 mr$$

or

$$r = \frac{L^2}{e^2m}$$
(16)

Finally we can eliminate the variable r in favor of the angular momentum L in our formula for the electron's total energy. We get

total energy  
of the electron 
$$E = \frac{-e^2}{2r}$$
  

$$= \frac{-e^2}{2(L^2/e^2m)}$$

$$= \frac{-e^2}{2} \left(\frac{e^2m}{L^2}\right)$$

$$= \frac{-e^4m}{2L^2}$$
(17)

In the formula  $-e^4m/2L^2$  for the electron's energy, only the angular momentum L changes from one orbit to another. If the energy of the nth orbit is  $E_n$ , then there must be a corresponding value  $L_n$  for the angular momentum of the orbit. Thus we should write



Angular momentum of a particle moving in a circle of radius r.

At this point, Bohr had the clue as to how to modify Newtonian mechanics in order to get his allowed energy levels. Suppose that angular momentum is quantized in units of some quantity we will call  $L_0$ . In the smallest orbit, suppose it has one unit, i.e.,  $L_1 = 1 \times L_0$ . In the second orbit assume it has twice as much angular momentum,  $L_2 = 2 L_0$ . In the nth orbit it would have n units

$$L_{n} = nL_{0} \qquad \begin{array}{c} quantization \\ of angular \\ momentum \end{array}$$
(19)

Substituting Equation 19 into Equation 18 gives

$$E_{n} = -\left(\frac{e^{4}m}{2L_{0}^{2}}\right)\frac{1}{n^{2}}$$
(20)

as the total energy of an electron with n units of angular momentum. Comparing Equation 20 with Bohr's energy level formula

$$E_n = -13.6 \text{ eV}\left(\frac{1}{n^2}\right)$$
 (10 repeated)

we see that we can explain the energy levels by assuming that the electron in the nth energy level has n units of quantized angular momentum  $L_0$ . We can also evaluate the size of  $L_0$  by equating the constant factors in Equations 10 and 20. We get

$$\frac{e^4m}{2L_0^2} = 13.6 \,\text{eV}$$
(21)

Converting 13.6 eV to ergs, and solving for  $L_0$  gives

$$\frac{e^4m}{2L_0^2} = 13.6 \text{ eV} \times 1.6 \times 10^{-12} \frac{\text{ergs}}{\text{eV}}$$

With  $e = 4.8 \times 10^{-10}$ esu and  $m = .911 \times 10^{-27}$ gm in CGS units, we get

$$L_0 = 1.05 \times 10^{-27} \frac{\text{gm cm}^2}{\text{sec}}$$
(22)

which turns out to be Planck's constant divided by  $2\pi$ .

$$L_0 = \frac{h}{2\pi} = \frac{6.63 \times 10^{-27} \text{gm cm/sec}}{2\pi}$$
  
= 1.05 × 10<sup>-27</sup>  $\frac{\text{gm cm}}{\text{sec}}$ 

This quantity, Planck's constant divided by  $2\pi$ , appears so often in physics and chemistry that it is given the special name "*h bar*" and is written  $\hbar$ 

$$\hbar \equiv \frac{h}{2\pi} \quad "h \, bar" \tag{23}$$

Using  $\hbar$  for  $L_0$  in the formula for  $E_n$ , we get Bohr's formula

$$E_n = -\left(\frac{e^4m}{2\hbar^2}\right)\frac{1}{n^2}$$
(24)

where  $e^4m/2\hbar^2$ , expressed in electron volts, is 13.6 eV. This quantity is known as the *Rydberg constant*. [*Remember that we are using CGS units, where e is in esu, m in grams, and h is erg-sec.*]

#### **Exercise 6**

Use Equation 21 to evaluate L<sub>0</sub>.

#### Exercise 7

What is the formula for the first Bohr radius in terms of the electron mass m, charge e, and Planck's constant  $\hbar$ . Evaluate your result and show that  $r_1 = .51 \times 10^{-8} \text{cm} = .51 \text{\AA}$ . (Answer:  $r_1 = \hbar^2/e^2 \text{m.}$ )

## **Exercise 8**

Starting from Newtonian mechanics and the Coulomb force law  $F = e^2/r^2$ , write out a clear and concise derivation of the formula

$$E_n = \frac{-e^4m}{2\hbar^2} \frac{1}{n^2}$$

Explain the crucial steps of the derivation.

A day or so later, on an empty piece of paper and a clean desk, see if you can repeat the derivation without looking at notes. When you can, you have a secure knowledge of the Bohr theory.

## **Exercise 9**

An ionized helium atom consists of a single electron orbiting a nucleus containing two protons as shown in Figure (8). Thus the Coulomb force on the electron has a magnitude

 $|F_{e}| = \frac{(e)(2e)}{r^{2}} = \frac{2e^{2}}{r^{2}}$ 

Figure 8 Ionized helium has a nucleus with two protons, surrounded by one electron.



a) Using Newtonian mechanics, calculate the total energy of the electron. (Your answer should be  $-e^2/r$ . Note that the r is not squared.)

b) Express this energy in terms of the electron's angular momentum L. (First calculate L in terms of r, solve for r, and substitute as we did in going from Equations 16 to 17.)

c) Find the formula for the energy levels of the electron in ionized helium, assuming that the electron's angular momentum is quantized in units of ħ.

d) Figure out whether ionized helium emits any visible spectral lines (lines with photon energies between 1.8 eV and 3.1 eV.) How many visible lines are there and what are their wavelengths?)

## **Exercise 10**

You can handle all single electron atoms in one calculation by assuming that there are z protons in the nucleus. (z = 1 for hydrogen, z = 2 for ionized helium, z = 3 for doubly ionized lithium, etc.) Repeat parts a), b), and c) of Exercise 9 for a single electron atom with z protons in the nucleus. (There is no simple formula for multi electron atoms because of the repulsive force between the electrons.)

## **DE BROGLIE'S HYPOTHESIS**

Despite its spectacular success describing the spectra of hydrogen and other one-electron atoms, Bohr's theory represented more of a problem than a solution. It worked only for one electron atoms, and it pointed to an explicit failure of Newtonian mechanics. The idea of correcting Newtonian mechanics by requiring the angular momentum of the electron be quantized in units of h, while successful, represented a bandaid treatment. It simply covered a deeper wound in the theory. For two centuries Newtonian mechanics had represented a complete, consistent scheme, applicable without exception. Special relativity did not harm the integrity of Newtonian mechanics-relativistic Newtonian mechanics is a consistent theory compatible with the principle of relativity. Even general relativity, with its concepts of curved space, left Newtonian mechanics intact, and consistent, in a slightly altered form.

The framework of Newtonian mechanics could not be altered to include the concept of quantized angular momentum. Bohr, Sommerfield, and others tried during the decade following the introduction of Bohr's model, but there was little success.

In Paris, in 1923, a graduate student Louis de Broglie, had an idea. He noted that light had a wave nature, seen in the 2-slit experiment and Maxwell's theory, and a particle nature seen in Einstein's explanation of the photoelectric effect. Physicists could not explain how light could behave as a particle in some experiments, and a wave in others. This problem seemed so incongruous that it was put on the back burner, more or less ignored for nearly 20 years.

De Broglie's idea was that, if light can have both a particle and a wave nature, perhaps electrons can too! Perhaps the quantization of the angular momentum of an electron in the hydrogen atom was due to the wave nature of the electron.

The main question de Broglie had to answer was how do you determine the wavelength of an electron wave?

An analogy with photons might help. There is, however, a significant difference between electrons and photons. Electrons have a rest mass energy and photons do not, thus there can be no direct analogy between the total energies of the two particles. But both particles have mass and carry linear momentum, and the amount of momentum can vary from zero on up for both particles. Thus photons and electrons could have similar formulas for linear momentum.

Back in Equation 34-13 we saw that the linear momentum p of a photon was related to its wavelength  $\lambda$  by the simple equation

$$\lambda = \frac{h}{p} \qquad \begin{pmatrix} de Broglie \\ wavelength \end{pmatrix}$$
(34-13)

De Broglie assumed that this same relationship also applied to electrons. An electron with a linear momentum p would have a wavelength  $\lambda = h/p$ . This is now called the *de Broglie wavelength*. This relationship applies not only to photons and electrons, but as far as we know, to all particles!

With a formula for the electron wavelength, de Broglie was able to construct a simple model explaining the quantization of angular momentum in the hydrogen atom. In de Broglie's model, one pictures an electron wave chasing itself around a circle in the hydrogen atom. If the circumference of the circle,  $2\pi r$  did not have an exact integral number of wavelengths, then the wave, after going around many times, would eventually cancel itself out as illustrated in Figure (9).

- Contrast indication of the i

**Figure 9** De Broglie picture of an electron wave cancelling itself out.



**Figure 10** If the circumference of the orbit is an integer number of wavelengths, the electron wave will go around without any cancellation.

But if the circumference of the circle were an exact integral number of wavelengths as illustrated in Figure (10), there would be no cancellation. This would therefore be one of Bohr's allowed orbits shown in Figure (6).

Suppose (n) wavelengths fit around a particular circle of radius  $r_n$ . Then we have

$$n\lambda = 2\pi r_n \tag{25}$$

Using the de Broglie formula  $\lambda = h/p$  for the electron wavelength, we get

$$n\left(\frac{h}{p}\right) = 2\pi r_n \tag{26}$$

Multiplying both sides by p and dividing through by  $2\pi$  gives

$$n\frac{h}{2\pi} = pr_n \tag{27}$$

Now  $h/2\pi$  is just  $\hbar$ , and  $pr_n$  is the angular momentum  $L_n$  (momentum times lever arm) of the electron. Thus Equation 27 gives

$$n\hbar = pr_n = L_n \tag{28}$$

Equation 28 tells us that for the allowed orbits, the orbits in which the electron wave does not cancel, the angular momentum comes in integer amounts of the angular momentum  $\hbar$ . The quantization of angular momentum is thus due to the wave nature of the electron, a concept completely foreign to Newtonian mechanics.



**Figure 10a--Movie** The standing waves on a circular metal band nicely illustrate de Broglie's waves

When a graduate student does a thesis project, typically the student does a lot of work under the supervision of a thesis advisor, and comes up with some new, hopefully verifiable, results. What do you do with a student that comes up with a strange idea, completely unverified, that can be explained in a few pages of algebra? Einstein happened to be passing through Paris in the summer of 1924 and was asked if de Broglie's thesis should be accepted. Although doubtful himself about a wave nature of the electron, Einstein recommended that the thesis be accepted, for de Broglie *just might be right*.

In 1925, two physicists at Bell Telephone Laboratories, C. J. Davisson and L. H. Germer were studying the surface of nickel by scattering electrons from the surface. The point of the research was to learn more about metal surfaces in order to improve the quality of switches used in telephone communication. After heating the metal target to remove an oxide layer that accumulated following a break in the vacuum line, they discovered that the electrons scattered differently. The metal had crystallized during the heating, and the peculiar scattering had occurred as a result of the crystallization. Davisson and Germer then prepared a target consisting of a single crystal, and studied the peculiar scattering phenomena extensively. Their apparatus is illustrated schematically in Figure (11), and their experimental results are shown in Figure (12). For their experiment, there was a marked peak in the scattering when the detector was located at an angle of  $50^{\circ}$  from the incident beam.

Davisson presented these results at a meeting in London in the summer of 1927. At that time there was a considerable discussion about de Broglie's hypothesis that electrons have a wave nature. Hearing of this idea, Davisson recognized the reason for the scattering peak. The atoms of the crystal were diffracting electron waves. The enhanced scattering at  $50^{\circ}$  was a diffraction peak, a maximum similar to the reflected maxima we saw back in Figure (33-19) when light goes through a diffraction grating. Davisson had the experimental evidence that de Broglie's idea about electron waves was correct after all.





#### Figure 12

Plot of intensity vs. angle for electrons scattered by a nickel crystal, as measured by Davisson and Germer. The peak in intensity at 50° was a diffraction peak like the ones produced by diffraction gratings. (The intensity is proportional to the distance out from the origin.)

## Chapter 36 Scattering of Waves

We will briefly interrupt our discussion of the hydrogen atom and study the scattering of waves by atoms. It was the scattering of electron waves from the surface of a nickel crystal that provided the first experimental evidence of the wave nature of electrons. Earlier experiments involving the scattering of x rays had begun to yield detailed information about the atomic structure of crystals.

Our main focus in this chapter will be an experiment developed in the early 1960s by Harry Meiners at R.P. I., that makes it easy for students to study electron waves and work with de Broglie's formula  $\lambda = h/p$ . The apparatus involves the scattering of electrons from a graphite crystal. The analysis of the resulting diffraction pattern requires nothing more than a combination of the de Broglie formula with the diffraction grating formula discussed in Chapter 33. We will use Meiner's experiment as our main demonstration of the wave nature of the electron.

## SCATTERING OF A WAVE BY A SMALL OBJECT

The first step in studying the scattering of waves by atoms is to see what happens when a wave strikes a small object, an object smaller in size than the wavelength of the wave. The result can be seen in the ripple tank photographs shown in Figure (1). In (1a), an incident wave is passing over a small object. You can see scattered waves emerging from the object. In (1b), the incident wave has passed, and you can see that the scattered waves are a series of circular waves, the same pattern you get when you drop a stone into a quiet pool of water. If the scattering object is smaller in size than the wavelength of the wave, as in Figure (1), the scattered waves contain essentially no information about the shape of the object. For this reason, you cannot study the structure of something that is much smaller than the wavelength of the wave you are using for the study. Optical microscopes, for example, cannot be used to study viruses, because most viruses are smaller than the wavelength of visible light. (Very clever work with optical microscopes allows one to see down to about 1/10th of the wavelength of visible light, to see objects like microtubules.)



a) Incident and scattered wave together.

## Figure 1

If the scattering object is smaller than a wavelength, we get circular scattered waves that contain little or no information about the shape of the object.



b) After the incident wave has passed.

## **REFLECTION OF LIGHT**

Using the picture of scattering provided by Figure (1), we can begin to understand the reflection of visible light from a smooth metal surface. Suppose we have a long wavelength wave impinging on a metal surface represented by a regular array of atoms, as illustrated in Figure (2). As the wave passes over the array of atoms, circular scattered waves emerge. As seen in Figure (2a), the scattered waves add up to produce a reflected wave coming back out of the surface. The angles labeled  $\theta_i$  and  $\theta_r$  in Figure (2b) are what are called the angle of incidence and angle of reflection, respectively. Since the scattered waves emerge at the same speed as the incident wave enters, it is clear from the geometry that the angle of incidence is equal to the angle of reflection. That is the main rule governing the reflection of light.

What happens inside the material depends upon details of the scattering process. Note that the reflected wavefront inside the material coincides with the incident wave. For a metal surface, the phases of the scattered waves are such that the reflected wave inside just cancels the incident wave and there is no wave inside. All the radiation is reflected. For other types of material that are not opaque, the incident and scattered waves do not cancel. Instead they add up to produce a new, transmitted wave whose crests move slower than the speed of light. This apparent slowing of the speed of light, due to the interference of transmitted and scattered waves, leads to the bending of a beam of light as it enters or leaves a transparent medium. *It is this bending that allows one to construct lenses and optical instruments*.

## Exercise 1

Using Figure (2), prove that the angle of incidence equals the angle of reflection.



When light reflects from a mirror, the angle of incidence equals the angle of reflection.



#### Figure 2a

A reflected wave is produced when the incident wave is scattered by many atoms. From this diagram, you can see why the angle of incidence equals the angle of reflection.

## **X RAY DIFFRACTION**

If the wavelength of the light striking a crystal becomes comparable to the spacing between atoms, we get a new effect. The scattered waves from adjacent atoms begin to interfere with each other and we get diffraction patterns.

The spacing between atoms in a crystal is of the order of a few angstroms. (An angstrom, abbreviated Å, is  $10^{-8}$ cm. An angstrom is essentially the diameter of a hydrogen atom.) Light with this wavelength is in the x ray region. Using Einstein's formula  $E = hf = hc/\lambda$ , but in the form

$$E(in eV) = \frac{12.4 \times 10^{-5} \text{ eV} \cdot \text{cm}}{\lambda (in cm)}$$

we see that photons with a wavelength of Å2 have an energy

$$E\begin{pmatrix}photon with 2\mathring{A}\\wavelength\end{pmatrix} = \frac{12.4 \times 10^{-5} \text{eV} \cdot \text{cm}}{2 \times 10^{-8} \text{cm}}$$
$$= 6,200 \text{ eV} \qquad (1)$$

This is a considerably greater energy than the 2 to 3 eV of visible photon.



**Figure 3** *Planes of atoms act like mirrors reflecting X rays.* 

When a beam of x rays is sent through a crystal structure, the x rays will reflect from the planes of atoms within the crystal. The process, called *Bragg reflection*, is illustrated for the example of a cubic lattice in Figure (3). The dotted lines connect lines of atoms, which are actually planes of atoms if you consider the depth of the crystal. An incident wave coming into the crystal can be reflected at various angles by various planes, with the angle of incidence equal to the angle of reflection in each case.

When the wavelength of the incident radiation is comparable to the spacing between atoms, we get a strong reflected beam when the reflected waves from one plane of atoms are an integral number of wavelengths behind the reflected waves from the plane above as illustrated in Figure (4). If it is an exact integral wavelength, then the reflected light from all the parallel planes will interfere constructively giving us an intense reflected wave. If, instead, there is a slight mismatch, then light from relatively distant planes will cancel in pairs and we will not get constructive interference. The argument is similar to the one used to find the maxima in a diffraction grating.



#### Figure 4

When the incident X ray wavelength equals the spacing between one of the sets of planes, the reflected waves add up to produce a maxima.

Thus with Bragg reflection you get an intense reflection only from planes of atoms, and only if the wavelength of the x ray is just right to produce the constructive interference described above. As a result, if you send an x ray beam through a crystal, you get diffraction pattern consisting of a series of dots surrounding the central beam, like those seen in Figure (5). Figure (5a) is a sketch of the setup and (5b) the resulting diffraction pattern for x rays passing through a silver bromide crystal whose structure is shown in (5c).



a) An incident beam of X rays is diffracted by the atoms of the crystal.



b) X ray diffraction pattern produced by a silver bromide crystal. (Photograph courtesy of R. W. Christy.)



## Figure 5

X ray diffraction study of a silver bromide crystal.

The main use of x ray diffraction has been to determine the structure of crystals. From the location of the dots in the x rays' diffraction photograph, and a knowledge of the wavelength of the x rays, you can figure out the orientation of and spacing between the planes of atoms. By using various wavelength x rays, striking the crystal at different angles, it is possible to decipher complex crystal structures. Figure (6) is one of many x ray diffraction photographs taken by J. C. Kendrew of a crystalline form of myoglobin. Kendrew used these x ray diffraction pictures to determine the structure of the myoglobin molecule shown in Figure (17-3). Kendrew was awarded the 1962 Nobel prize in chemistry for this work.



#### Figure 6

One of the X ray diffraction photographs used by Kendrew to determine the structure of the Myoglobin molecule.



**Figure 17-3** The Myoglobin molecule, whose structure was determined by X ray diffraction studies.

## **Diffraction by Thin Crystals**

The diffraction of waves passing through relatively thin crystals can also be analyzed using the diffraction grating concepts discussed in Chapter 33. Suppose for example, we had a thin crystal consisting of a rectangular array of atoms as shown in Figure (7a). The edge view of the array is shown in (7b). Here each dot represents the end view of a line of atoms.



#### **Figure 7a**

Front view of a rectangular array of atoms in a thin crystal.



#### Figure 7b

Edge view with an incident wave. Each dot now represents one of the line of atoms in Figure (7a).



#### Figure 8

The waves emerging from a diffraction grating have a similar structure as waves scattered by a line of atoms.

Now suppose a beam of waves is impinging upon the crystal as indicated in Figure (7b). The impinging waves will scatter from the lines of atoms, producing an array of circular waves as shown.

Compare this with Figure (8), a sketch of waves emerging from a diffraction grating. The scattered waves from the lines of atoms, and the waves emerging from the narrow slits have a similar structure and therefore should produce similar diffraction patterns.



Various lines of atoms can imitate slits in a diffraction grating.

#### Figure 10a

A laser beam sent through a single grating. The lines of the grating were 25 microns wide, spaced 150 microns apart.





There is one major difference between the array of atoms in Figure (7) and the diffraction grating of Figure (8). In the crystal structure there are numerous sets of lines of atoms, some of which are indicated in Figure (9). Each of these sets of lines of atoms should act as an independent diffraction grating, producing its own diffraction pattern. The main sets of lines are horizontal and vertical, thus the main diffraction pattern we should see should look like that produced by two diffraction gratings crossed at right angles. Sending a laser beam through two crossed diffraction gratings produces the image shown in Figure (10). In Figure (10a), the laser beam is sent through a single grating. In (10b) we see the effect of adding another grating crossed at right angles.

#### Exercise 2

In Figure (10a) the maxima seen in the photograph are 1.68 cm apart and the distance from the grating to the screen is 4.00 meters. The wavelength of the laser beam is  $6.3 \times 10^{-5}$  cm. What is the spacing between the slits of the diffraction grating?

## Exercise 3

In Figure (11), a laser beam is sent through two crossed diffraction gratings of different spacing. Which image, (a) or (b) is oriented correctly? (What happens to the spacing of the maxima when you make the grating lines closer together?)

#### Figure 11

Two diffraction gratings with different spacing are crossed. As shown, the vertical lines are farther apart than the horizontal ones. Which of the two images of the resulting diffraction pattern has the correct orientation?

					L
_					
-			_		-
_	_	_	_	_	_
-					_





### Figure 10b

A laser beam sent through crossed diffraction gratings. Again the lines of the grating were 25 microns wide, spaced 150 microns apart.





## THE ELECTRON DIFFRACTION EXPERIMENT

One of the main differences between the scattering of x rays and of electrons is that x ray photons interact less strongly with atoms, with the result that x rays can penetrate deeply into matter. This enables doctors to photograph through flesh to observe broken bones, or engineers to photograph through metal looking for hidden flaws. Electrons interact strongly with atoms, do not penetrate nearly as deeply, and therefore are well suited for the study of the structure of surfaces or thin crystals where you get considerable scattering from a few layers of atoms.



## Figure 12

The hexagonal array of atoms in one layer of a graphite crystal. Lines of atoms in this crystal act as crossed diffraction gratings.

## The Graphite Crystal

Graphite makes an ideal substance to study by electron scattering because graphite crystals come in thin sheets. A graphite crystal consists of a series of planes of carbon atoms. Within one plane the atoms have the hexagonal structure shown in Figure (12), reminiscent of the tiles often seen on bathroom floors. The spacing between neighboring atoms in each hexagon is 1.42Å as indicated at the bottom of Figure (12).

The atoms within a plane are very tightly bound together. The hexagonal array forms a very strong framework. The planes themselves are stacked on top of each other at the considerable distance of 3.63 Å as indicated in Figure (13). The forces between these planes are weak, allowing the planes to easily slide over each other. The result is that graphite is a slippery substance, making an excellent dry lubricant. In contrast, the strength within a plane makes graphite an excellent strengthening agent for epoxy. The resulting carbon filament epoxies, used for constructing racing boat hulls, light airplanes and stayless sailboat masts, is one of the strongest plastics available.



#### Figure 13

Edge view of the graphite crystal, showing the planes of atoms. The planes can easily slide over each other, making the substance slippery.

## **The Electron Diffraction Tube**

The electron diffraction experiment where we sent a beam of electrons through a graphite crystal, can be viewed either as an experiment to demonstrate the wave nature of electrons or as an experiment to study the structure of a graphite crystal. Perhaps both.

The apparatus, shown in Figure (14), consists of an evacuated tube with an electron gun at one end, a graphite target in the middle, and a phosphor screen at the other end. A finely collimated electron beam can be aimed to strike an individual flake of graphite, producing a single crystal diffraction pattern on the phosphor screen. Usually you hit more than one crystal and get a multiple image on the screen, but with some adjustment you can usually obtain a single crystal image.





#### Figure 14

Electron diffraction apparatus. An electron beam, produced by an electron gun, strikes a graphite crystal located near the center of the evacuated tube. The original beam and the scattered electrons strike a phosphor screen located at the end of the tube.

## **Electron Wavelength**

The accelerating voltage required to produce a good diffraction pattern is in the range of 6,000 volts. As our first step in the analysis, let us use the de Broglie wavelength formula to calculate the wavelength of 6,000 eV electrons.

The rest energy of an electron is .51 MeV, or 510,000 eV, far greater than the 6,000 eV we are using in this experiment. Since the 6,000 eV kinetic energy is much less than the rest energy, we can use the nonrelativistic formula  $1/2 \text{ mv}^2$  for kinetic energy. First converting 6,000 eV to ergs, we can equate that to  $1/2 \text{ mv}^2$  to calculate the speed v of the electron. We get

6000 eV × 1.6 × 10<sup>-12</sup> 
$$\frac{\text{ergs}}{\text{eV}}$$
 = 1/2 m<sub>e</sub>v<sup>2</sup> (2)

With the electron mass  $m_e = .911 \times 10^{-27} gm$ , we get

$$v^{2} = \frac{2 \times 6000 \times 1.6 \times 10^{-12} \text{ergs}}{.911 \times 10^{-27} \text{gm}}$$
$$= 21.1 \times 10^{18} \frac{\text{cm}^{2}}{\text{sec}^{2}}$$

$$v = 4.59 \times 10^9 \text{ cm/sec}$$
 (3)

which is slightly greater than 10% the speed of light.

The next step is to calculate the momentum of the electron for use in de Broglie's formula. We have

$$p = mv$$
  
= .911 × 10<sup>-27</sup>gm × 4.59 × 10<sup>9</sup> cm/sec (4)  
= 4.18 × 10<sup>-18</sup> gm cm/sec

Finally using de Broglie's formula we have

$$\lambda = \frac{h}{p} = \frac{6.63 \times 10^{-27} \text{gm cm}^2/\text{sec}}{4.18 \times 10^{-18} \text{gm cm/sec}}$$
$$\lambda_{\text{electron}} = 1.59 \times 10^{-9} \text{cm} = .159 \text{ Å}$$
(5)

Thus the wavelength of the electrons we are using in this experiment is about one tenth the spacing between atoms in the hexagonal array.

## **Exercise 4**

Calculate the wavelength of a 6000 eV photon. What would cause such a difference in the wavelengths of a photon and an electron of the same energy?



**Figure 15a** Single grating diffraction pattern.



**Figure 15b** *Two grating diffraction pattern.* 



**Figure 15c** Diffraction pattern from three crossed gratings.









**Figure 12 (section)** Three sets of lines of atoms act as three crossed diffraction gratings with 2.13 angstrom spacing.

## **The Diffraction Pattern**

What should we see when a beam of waves is diffracted by the hexagonal array of atoms in a graphite crystal? Looking back at the drawing of the graphite crystal, Figure (12), we see that there are prominent sets of lines of atoms in the hexagonal array. To make an effective diffraction grating, the lines of atoms have to be equally spaced. We have marked three sets of equally-spaced lines of atoms, each set being at an angle of  $60^{\circ}$  from each other. We expect that these lines of atoms should produce a diffraction pattern similar to three crossed diffraction gratings.

In Figure (15), we are looking at the diffraction we get when a laser beam is sent through three crossed diffraction gratings. In (15a), we have 1 diffraction grating. In (15b) a second grating at an angle of  $60^{\circ}$  has been added. In (15c) we have all three gratings, and see a hexagonal array of dots surrounding the central beam, the central maximum.

Figure (16) is the electron diffraction pattern photographed from the face of the electron diffraction tube shown in Figure (14). We clearly see an hexagonal array of dots expected from our diffraction grating analysis. On the photograph we have superimposed a centimeter scale so that measurements may be made from this photograph.



#### Figure 16

Diffraction pattern produced by a beam of electrons passing through a single graphite crystal. The energy of the electrons was 6000 eV.

The electron diffraction apparatus allows us to move the beam around, so that we can hit different parts of the target. In Figure (16), we have essentially hit a single crystal. When the electron beam strikes several graphite crystals at the same time, we get the more complex pattern seen in Figure (17).

## Analysis of the Diffraction Pattern

Let us begin our analysis of the diffraction pattern by selecting one set of dots in the pattern that would be produced by one set of lines of atoms in the crystal. The dots and the corresponding lines of atoms are shown in Figure (18). In (18a) we see that the spacing  $Y_{max}$  between the dots on the screen is 1.33 cm. These horizontal dots correspond to the maxima for a vertical set of lines of atoms indicated in (18c). In (18b) we are reminded that the distance from the target to the screen is 18 cm. Using the diffraction grating formula, we can calculate the wavelength of the electron waves that produce this set of maxima.



Figure 17

Diffraction pattern produced by a beam of electrons passing through multiple graphite crystals.

Using the diffraction formula, Equation 33-3, and noting that  $Y_{max} \ll D$ , we have

$$\lambda = Y_{max} \frac{d}{\sqrt{D^2 + Y_{max}^2}} \approx Y_{max} \frac{d}{D}$$
$$= \frac{1.33 \text{ cm} \times 2.13 \times 10^{-8} \text{ cm}}{18 \text{ cm}}$$
$$\lambda = 1.57 \times 10^{-9} \text{ cm} \qquad (6)$$

which agrees well with Equation 5, the calculation of the electron wavelength using the de Broglie wavelength formula.





The diffraction grating maxima from one set of lines in the graphite crystal. You can see that  $3y_{max} = 4cm$ , so that  $y_{max} = 1.33$  cm.



**Figure 18b** *Top view of the electron diffraction apparatus..* 



**Figure 18c** The vertical lines of atoms in the graphite crystal that produce the horizontal row of dots seen in (a).



#### Figure 19

It is easy to find a second set of effective gratings, rotated  $30^{\circ}$  from the first set, and with a narrower spacing.

#### Figure 20

We have highlighted the maxima produced by this set of lines. Note that the more narrowly spaced lines produce more widely spaced maxima.





## **Other Sets of Lines**

With a careful analysis of the lines of atoms in the hexagonal ray of atoms, one can explain all the dots of the diffraction pattern of Figure (16). For example, in Figure (19) we see that there is another set of lines that are rotated at an angle of  $30^{\circ}$  and more closely spaced than our original set. In Figure (20), we have highlighted a set of dots in the diffraction pattern that are rotated by an angle of  $30^{\circ}$  and more widely spaced than the dots we have been analyzing. Since more closely spaced maxima, we should suspect that the highlighted maxima result from this new set of lines. The point of Exercise 5 is to see if this is true.

## Exercise 5

(a) Explain why more closely spaced atoms should produce more widely spaced dots in the diffraction pattern.

(b) Assuming that the dots highlighted in Figure (20) are produced by the lines of atoms shown by dotted lines in set 2 of the effective gratings, calculate the wavelength of the waves producing the dots. Compare your results with our previous analysis.

## **Exercise 6**

Suppose that a beam of neutrons rather than electrons were fired at the graphite crystal. Assuming that neutrons also obey the de Broglie relationship  $\lambda = h/p$ , what should be the kinetic energy, in eV, of the neutrons in order to produce the same diffraction pattern with the same spacing between dots?

## **Student Projects**

The crossed diffraction gratings used to obtain the various laser diffraction patterns in this chapter, were created using the Adobe Illustrator program, and then printed on film using a Linatronic imagesetter at a local desktop publishing company. The one micron resolution of the imagesetter allowed us to construct various grating and dot patterns that produced reasonable diffraction patterns with a laser.



## Figure 21

Microphotograph of the three crossed diffraction gratings. The lines are 25 microns wide and 100 microns apart, on centers. (Student project by Brady Beale and Amy Coughlin.)

Several students doing project work with these gratings and dot patterns suspected that some patterns were not as good as they should be and took microscope photographs of them. They found that lines or dots as small as 10 microns wide tended to be filled in and blotchy, but lines or dots 25 microns wide came out fairly well as can be seen in Figures (21) and (22). Figure (23) is the laser diffraction pattern produced by a laser beam passing through the hexagonal dot pattern of Figure (22).



## Figure 22

Microphotograph of a hexagonal dot pattern. The dots are 25 microns in diameter and 100 microns apart. (Student project by Brady Beale and Amy Coughlin.)



**Figure 15c** Diffraction pattern produced by a laser beam going through the three crossed gratings of Figure 21.



**Figure 23** Diffraction pattern produced by a laser beam going through the hexagonal dot pattern of Figure 22.

## **Student project by Gwendylin Chen**

In our discussion of the diffraction of waves by the atoms of a crystal, we pointed out that waves should emerge from a line of atoms in much the same wah that they do from the slits of a diffraction grating. The two situations were illustrated in Figures (7b and 8) reproduced below.

That a slit and a line produce similar diffraction patterns was clearly illustrated in a project by Gwendylin Chen. While working with a laser, she observed that when the beam passed over a strand of hair it produced a single slit diffraction pattern superimposed on the image of the beam itself. Here we have reproduced Gwendylin's experiment. Figure (24) is a photograph of a slit made from two scapel blades, and a strand of Gwendylin's hair. We tried to make the width of the slit the same as the width of the hair. The two circles indicate where we aimed the laser for the two diffraction patterns.

The results are seen in Figure (25). The diffraction patterns are almost identical. The only difference is that when the beam passes over the hair, it continues on landing in the center of the diffraction pattern.



Figure 7b Edge view of a thin crystal with an incident wave. Each dot now represents one of the line of atoms in the crystal.



Figure 8

The waves emerging from a diffraction grating have a structure similar to the waves scattered by a line of atoms.

## slit formed by two scapels





Slit and hair used to produce diffractiopn patterns. The circles indicate where we aimed the laser.



a) single slit diffraction pattern



b) diffraction pattern produced by strand of hair

**Figure 25** *Comparason of diffeaction patterns.*  incident laser beam

## Chapter 37 Lasers, a Model Atom and Zero Point Energy

"Once at the end of a colloquium I heard Debye saying something like: 'Schrödinger, you are not working right now on very important problems...why don't you tell us some time about that thesis of de Broglie, which seems to have attracted some attention?' So in one of the next colloquia, Schrödinger gave a beautifully clear account of how de Broglie associated a wave with a particle, and how he could obtain the quantization rules ... by demanding that an integer number of waves should be fitted along a stationary orbit. When he had finished, Debye casually remarked that he thought this way of talking was rather childish ... To deal properly with waves, one had to have a wave equation."

FELIX BLOCK, in an address to the American Physical Society in 1976.

Schrödinger took Debye's advice, and in the following months devised a wave equation for the electron wave, an equation from which one could calculate the electron energy levels. The structure of the hydrogen atom was a prediction of the equation without arbitrary assumptions like those needed for the Bohr theory. The wave nature of the electron turned out to be the key to the new mechanics that was to replace Newtonian mechanics as the fundamental theory.

In the next chapter we will take a look at some of the electron wave patterns determined by Schrödinger's equation, and see how these patterns, when combined with the Pauli exclusion principle and the concept of electron spin, begin to explain the chemical properties of atoms and the structure of the periodic table.

The problem one encounters when discussing the application of Schrödinger's equation to the hydrogen atom, is that relatively complex mathematical steps are required in order to obtain the solutions. These steps are usually beyond the mathematical level of most introductory physics and chemistry texts, with the result that students must simply be shown the solutions without being told how to get them. We will have to do the same in the next chapter.

In this chapter we will study a model atom, one in which we can see how the particle-wave nature of the electron leads directly to quantized energy levels and atomic spectra. The basic idea, which we illustrate with the model atom, is that whenever you have a wave confined to some region of space, there will be a set of allowed standing wave patterns for that wave. Whether the patterns are complex or simple depends upon the way the wave is confined. If the wave is also a particle, like an electron or photon, you can then use the particle wave nature to calculate the energy of the particle in each of the allowed standing wave patterns. These energy values are the quantized energy levels of the particle.

An example of a set of simple standing waves that are easily analyzed is found in the laser. It is essentially the laser standing wave patterns that we use for our model atom. For this reason we begin the chapter with a discussion of the laser and how the photon standing waves are established. In the model atom the photon standing waves of the laser are replaced by electron standing waves.

An analysis of the model atom shows why any particle, when confined to some region of space, must have a non zero kinetic energy. The smaller the region of space, the greater this so called **zero point** kinetic energy. When these ideas are applied to the atoms in liquid helium, we see why helium does not freeze even at absolute zero. We also see why the entropy definition of temperature must be used at these low temperatures.

## THE LASER AND STANDING LIGHT WAVES

The laser, the device that is at the heart of your CD player and fiber optics communications, provides a common example of a standing light wave. In most cases a laser consists of two parallel mirrors with standing light waves trapped between the mirrors as illustrated in Figure (1). The light comes from radiation emitted by excited atoms that are located within the standing wave.

How the light radiated by the excited atoms ends up in a standing wave is a story in itself. An atom excited to a high energy level can drop down to a lower level by emitting a photon whose energy is the difference in energy of the two levels. This photon will have the wavelength of the spectral line associated with those two levels.

Spectral lines are not absolutely sharp. For example, due to the Doppler effect, thermal motion slightly shifts the wavelength of the emitted radiation. If the atom is moving toward you when it radiates, the wavelength is shifted slightly towards the blue. If moving away, the shift is toward the red. In addition the photons are radiated in all directions, and waves from different photons have different phases. Even in a sharp spectral line the light is a jumble of directions and phases, giving what is called *incoherent* light.

In contrast the light in a laser beam travels in one direction, the phases of the waves are lined up and there is almost no spread in photon energies. This is the beam of *coherent* light which made it so easy for us to study interference effects like those we saw in the two slit and multiple slit diffraction patterns. These patterns would be much more difficult to observe if we had to use incoherent light.



## Figure 1

Laser consisting of two parallel mirrors with standing light waves trapped between the mirrors

The purity of the light in a laser beam depends upon the standing light wave pattern created by the two mirrors, and upon a quantum mechanical effect discovered by Einstein in 1915.

Einstein found that there were two distinct ways an excited atom could radiate light, either by *spontaneous emission* or *stimulated emission*. An example of spontaneous emission is when an excited atom is all by itself and eventually drops down to a lower energy level. The emitted photon can come out in any direction and can be Doppler shifted.

If, however, a photon with the right energy passes by the excited atom, there is some chance that the atom will emit a photon *exactly* like the one passing by. This is called stimulated emission. (The energy of the passing photon has to be close to the energy the atom would naturally radiate.)

It is the process of stimulated emission that can lead to a laser beam. Suppose we have a gas of excited atoms located between parallel mirrors. At first the atoms radiate spontaneously in all directions. (We assume that there is some mechanism to excite the atoms). After a while one of the photons hits a mirror straight on and starts reflecting back and forth between the parallel mirrors. As the photon moves back and forth, it passes by an excited atom, stimulating that atom to emit an identical photon.

Now there are two identical photons bouncing back and forth. Each is likely to stimulate another atom to emit an identical photon, and we have four identical photons, etc. Soon there are so many identical photons moving through the excited atoms that there is little chance that an atom can radiate spontaneously. All the radiation is stimulated and all the photons are identical to the one that started bouncing back and forth between the mirrors.

The mirrors on the ends of the laser are not perfect reflectors, a few percent of the photons striking the mirror pass through, forming the beam produced by the laser. The photons lost to the laser beam are continually replaced by new identical photons being emitted by stimulated emission. One of the tricky technical parts of constructing a laser is to maintain a continuous supply of excited atoms. There are various ways of doing this that we need not discuss here.

## **Photon Standing Waves**

The photons bouncing back and forth between the mirrors in a laser are in an allowed standing wave pattern. Back in Chapter 15 in our discussion of standing waves on a guitar string, we saw that only certain standing wave patterns were allowed, those shown in Figure (15-15) reproduced here which had an integral or half integral number of wavelengths between the ends of the string. For photons trapped between two mirrors, the allowed standing wave patterns are also those with an integral or half integral number of wavelengths between the mirrors, as indicated schematically in Figure (2).



Figure 15-15 (reproduced)

On a guitar string only certain standing wave patterns which have an integral or half intergral number of wavelengths between the ends of the string are allowed. Because of the simple geometry, we do not need to solve a wave equation to determine the shape of these standing light waves. The waves are sinusoidal, and the allowed wavelength are given by the same formula as for the allowed waves on a guitar string, namely

$$\lambda_{n} = \frac{2D}{n} \qquad wavelength of the \\ nth standing wave$$
(1)

where D is the separation between the mirrors.





## **Photon Energy Levels**

The special feature of the standing light wave is that the light has both a wave and a particle nature. Equation 1, which tells us the allowed wavelengths, is all we need to know about the wave nature of the light. The particle nature is described by Einstein's photoelectric effect formula  $E = hf = hc/\lambda$ . Applying this formula to the photons in the standing wave, we find that a photon with an allowed wavelength  $\lambda_n$  has a corresponding energy  $E_n$  given by

$$E_n = \frac{hc}{\lambda_n}$$
(2)

Because only certain wavelengths  $\lambda_n$  are allowed, only certain energy photons, those with an energy  $E_n$ are allowed between the mirrors. We can say that the photon energies are quantized. If the separation of the



#### Figure 3

Energy level diagram for a photon trapped between two mirrors.

mirrors is D, then from Equation 1 ( $\lambda_n = 2D/n$ ), and Equation 2 ( $E_n = hc/\lambda_n$ ), we find that the quantized values of  $E_n$  are

$$E_{n} = \frac{hc}{\lambda_{n}} = \frac{hc}{\frac{2D}{n}}$$

$$E_{n} = n\frac{hc}{2D}$$
(3)

From Equation 3 we can construct an energy level diagram for the photons trapped between the mirrors. In contrast to the energy level diagram for the hydrogen atom, the photon energies start at zero because there is no potential energy. We see that the levels are equally spaced, a distance hc/2D apart.

#### **Exercise 1**

If you could have two mirrors 1Å apart (the size of a hydrogen atom) what would be the energy, in eV, of the lowest 5 energy levels for a photon trapped between the mirrors?

## A MODEL ATOM

Now imagine that we replace the photons trapped between two mirrors with an electron between parallel walls located a distance D apart, as shown in Figure (4). For this model, the allowed standing wave patterns are again similar to the guitar string standing waves. The allowed electron wavelengths are

$$\lambda_{n} = \frac{2D}{n} \qquad \begin{array}{l} allowed wavelength \\ of an electron trapped \\ between two walls \end{array}$$
(1a)

The difference between having a photon trapped between mirrors and an electron between walls, is the formula for the energy of the particle. If the energy of the electron is non relativistic, then the formula for its



**Figure 4** Electron trapped between two walls

kinetic energy is  $1/2 \text{ mv}^2$ , not the Einstein formula  $\text{E} = \text{hc}/\lambda$  that applies to photons. The difference arises because the electron has a rest mass while the photon does not.

For the electron trapped between walls, there is no electric potential energy like there was in the hydrogen atom. Thus we can take  $1/2 \text{ mv}^2$  as the formula for the electron's total energy, ignoring the electron's rest mass energy as we usually do in non relativistic calculations.

To relate the kinetic energy to the electron's allowed wavelength  $\lambda_n$ , we use de Broglie's formula  $p = h/\lambda$ . The easy way to do this is to express the energy  $1/2 \text{ mv}^2$  in terms of the electron's momentum p = mv. We get

$$E = 1/2 \text{ mv}^2 = \frac{1}{2m} (\text{mv})^2$$
$$E = \frac{p^2}{2m}$$
(4)

Next use the de Broglie formula  $p = h/\lambda$  to give us

$$E_n = \frac{\left(h/\lambda_n\right)^2}{2m} = \frac{h^2}{2m\lambda_n^2}$$
(5)

as the formula for the energy of an electron of wavelength  $\lambda_n$  .

Finally use Equation 1,  $\lambda_n$  = 2D/n , for allowed electron wavelengths to get

$$E_{n} = \frac{h^{2}}{2m\left(\frac{2D}{n}\right)^{2}}$$

$$E_{n} = n^{2}\left(\frac{h^{2}}{8mD^{2}}\right)$$
(6)

This is our equation for the energy levels of an electron trapped between two plates separated by a distance D. The corresponding energy level diagram is shown in Figure (5). The energy levels go up as  $n^2$  instead of being equally spaced as they were in the case of a photon trapped between two mirrors.

If the electron is in one of the higher levels and falls to a lower one, it will get rid of its energy by emitting a photon whose energy is equal to the difference in the energy of the two levels. Thus the trapped electron should emit a spectrum of radiation with sharp spectral lines, where the lines correspond to energy jumps between levels just as in the hydrogen atom. Thus the electron trapped between plates is effectively a model atom, complete with an energy level diagram and spectral lines.



electron energy levels

#### Figure 5

Energy level diagram for an electron trapped between two walls.

Our model atom is not just a fantasy. With the techniques used to fabricate microchips, it has been possible to construct tiny boxes, the order of a few angstroms across, and trap electrons inside. An electron microscope photograph of these *quantum dots* as they are called, is shown in Figure (6). The allowed standing wave patterns are reasonably well represented by the sine wave patterns of Figure (5), where D is the smallest dimension of the box. Thus we predict that electrons trapped in these boxes should have allowed energies  $E_n$  close to those given by Equation (6), and emit discrete line spectra like an atom. This is precisely what they do. (Some of the low energy jumps are shown in Figure 7.)

In calculating with the model atom we have not fudged in any way by modifying Newtonian mechanics or even picturing a wave chasing itself around in a circle. We see a spectrum resulting purely from a combination of the particle nature and the wave nature of electrons and photons, where the connection between the two points of view is de Broglie's formula  $p = h/\lambda$ .



#### Figure 6

Grid of quantum dots. These cells are made on a silicon wafer with the same technology used in making electronic chips. An electron trapped in one of these cells has energy levels similar to those of our model atom. (See Scientific American, Jan. 1993, p118.)

## Exercise 2

Assume that an electron is trapped between two walls a distance D apart. The distance D has been adjusted so that the lowest energy level is  $E_1 = 0.375 \text{ eV}$ .

(a) What is D?

(b) What are the energies, in eV, of the photons in the six longest wavelength spectral lines radiated by this system? Draw the energy level diagram for this system and show the electron jumps corresponding to each spectral line.

(c) What are the corresponding wavelengths, in cm, of these six spectral lines?

(d) Where in the electromagnetic spectrum (infra red, visible, or ultra violet) do each of these spectral lines lie? If any of these lines are visible, what color are they? (Partial answer: the photon energies are 1.125, 1.875, 2.625, 3.00, 3.375, and 4.125 eV)

## **Exercise 3**

Explain why an electron, confined in a box, cannot sit at rest. This is an important result whose consequences will be discussed next. Try to answer it now.



## Figure 7

When an electron falls from one energy level to another, the energy of the photon it emits equals the energy lost by the photon.

## ZERO POINT ENERGY

One of the immediate consequences of the particlewave nature of the electron is that a confined electron can never be at rest. The smaller the confinement, the greater the kinetic energy the electron must have. This follows from the fact that at least half a wavelength of the electron's wave must fit within the confining region. If D is the length of the smallest dimension of the confining region, then the electron's wavelength cannot be greater than 2D. But the smaller D is, the shorter the electron's wavelength is, the greater its kinetic energy.

The de Broglie wavelength formula  $\lambda = h/p$  applies not only to photons and electrons, but to any particle, even an entire atom. As a result, an atom confined to a region of size D should have a wavelength no greater than  $\lambda_1 = 2D$ , and thus a minimum kinetic energy

$$E_{\min} = \frac{h^2}{8m_{atom}D^2}$$
(7)

where we simply replaced the electron's mass by the atom' mass in Equation 7. Equation 7 is somewhat approximate if the atom is confined on all sides in a three dimensional box, but it is reasonably accurate if D is the smallest dimension of the box.

An atom in a solid or a liquid is an example of a particle confined in a box. The atom is confined by its neighboring atoms as illustrated in Figure (8). We may think of its neighbors as forming a box of size D where D is the average spacing between atoms. Thus atoms in solids or liquids have a minimum kinetic energy given by Equation 7, and the atoms must be in continual motion *no matter how low the temperature*! Cooling the solid cannot get rid of this so-called *zero point energy*.

#### Figure 8

A helium atom in liquid helium is confined by its neighbors. As a result it has a zero point energy like an electron confined between walls.



#### **Exercise 4**

In liquid helium, the helium atoms are about 3Å apart and the atoms have a mass essentially equal to 4 times the mass of a proton.

(a) what is the zero point energy, in ergs, of helium atoms in liquid helium?

(b) at what temperature T is the helium atom's thermal kinetic energy 3/2 kT equal to the zero point energy calculated in part (a)? [Answer: (a)  $9.1 \times 10^{-16}$ ergs, (b) 4.42 kelvin.]

Helium is an especially interesting substance to study at low temperatures because it is the only substance that remains a liquid all the way down to absolute zero. The only way you can freeze helium is to take it down to very low temperatures, and then squeeze it at relatively high pressure.

In all other substances, at low enough temperatures the atoms settle down to a solid array. To melt the solid, you have to add enough thermal energy to disrupt the molecular bonds that hold the atoms in a more or less fixed array.

Why can't helium atoms be cooled to the point where molecular forces dominate and the atoms form a solid array? Part of the answer is that the molecular forces between helium atoms are very weak, the weakest there is between any atoms. Consequently you have to go to very low temperatures before helium gas even becomes a liquid. At atmospheric pressure, helium becomes a liquid at 4.5 kelvins. To turn liquid helium into a solid you should have to go to still lower temperatures.

From Exercise 4, you saw that, in one sense, you cannot get helium to a lower temperature, at least as far as the kinetic energy of the atoms is concerned. The zero point energy of the atoms is as big as the thermal energy that the atoms would have at a few kelvin-- 4.4 kelvin by our rough estimate in Exercise 4. As a result, cooling the helium further cannot remove enough kinetic energy to allow the helium liquid to freeze. Helium thus remains a liquid all the way down to absolute zero.

## **Definition of Temperature**

This discussion raises interesting questions about the very concept of temperature. Our initial experimental definition of temperature was the ideal gas thermometer, which, as we saw from the derivation of the ideal gas law, is based on the thermal kinetic energy of the particles. The simple idea of absolute zero was the point where all the thermal kinetic energy was gone and the atoms were at rest. Now we see that no matter how much thermal kinetic energy" remains. This is not a problem at ordinary temperatures, but it can significantly affect the behavior of matter at temperatures close to absolute zero.

At low temperatures, the ideal gas thermometer is not adequate, and a new definition of temperature is needed. That new definition is provided by the efficiency of Carnot's heat engine. As we suggested in Chapter 17, this gives us a definition of temperature based, not on the kinetic energy of the molecules, but upon the degree of randomness or disorder. A system at absolute zero is as perfectly ordered as it can be. If zero point energy is required by the particle wave nature of the atoms, if it cannot be removed, then the most organized, least disordered state of the system must include this zero point energy. Helium can go to its most ordered state at absolute zero, retain its zero point energy, and remain a liquid.







**Figure 16-41 (reproduced)** *Standing waves on a drumhead.* 

## TWO DIMENSIONAL STANDING WAVES

In our discussion of percussion instruments in Chapter 16, we saw that a drumhead has a set of allowed standing wave patterns or normal modes, in some ways like the standing waves or normal modes on a guitar string. On a guitar string we have one dimensional waves, while the drumhead has the two dimensional wave patterns. The six lowest frequency patterns are shown in Figure (16-41) repeated here. We could excite and observe individual standing waves using the apparatus shown in Figure (16-40) also shown again here.

That we get the same kind of standing wave patterns on an atomic scale is seen in Figure (9), which is a recent tunneling microscope image of an electron standing wave on the surface of a copper crystal. The standing wave, which is formed inside a corral of 48 iron atoms, has the same shape as one of the allowed standing waves on a drumhead. (This particular standing wave pattern is excited because the average wavelength in the standing wave is closest to the wavelength of the conduction electrons at the surface of the copper.)

A colleague Geoff Nunes, who works with scanning microscopes, describes the image: "The incredible power of today's personal computers has been made possible by our ability to make smaller and smaller transistors. The smallest transistor one could imagine building would be made up of single atoms. In a dramatic series of experiments at IBM, Don Eigler and his co-workers have shown how to use a tunneling microscope to move and arrange single atoms."





"This picture (Figure 9) shows a ridge of 48 iron atoms arranged in a circle on the surface of a copper crystal. Electrons in the copper are reflected from these iron atoms much as the waves on the surface of a pond are reflected from anything at the surface: rocks, weeds, the shoreline. Inside the ring, the electron waves form a beautifully symmetric pattern. This pattern occurs often in the physical world. For example it is the shape that the head of a drum forms when struck. You can easily observe a similar pattern by gently skidding the base of a Styrofoam cup full of coffee across the surface of a table."



## Figure 16-40 (reproduced)

Exciting and observing the standing waves on a drumhead.



## Figure 9

Conduction electrons on the surface of a copper crystal, forming a standing wave inside a corral of 48 iron atoms. The shape is the same as one of the symmetric standing waves on a drumhead. (Photo credits: Crommie and Eigler/IBM.)

# Chapter 38 Atoms

The focus of this chapter will be on the allowed electron standing wave patterns in hydrogen, as calculated by Schrödinger's wave equation. We will not work with Schrödinger's equation itself, which involves derivatives in both space and time, and requires fairly advanced mathematical techniques to handle. But this is not a terrible loss, because the resulting wave patterns are well known, and are all we need in order to understand much of the structure and behavior of atoms.

As we saw at the end of the last chapter, when we go to two dimensions, the standing wave patterns become more complex. For example, to find the drumhead standing wave patterns, we either had to do an experiment to observe the patterns, or solve a wave equation to calculate them. To determine electron waves in hydrogen, our only option is to rely on Schrödinger's equation. The resulting standing waves are three dimensional in shape, and do not have sharp edges like the drumhead waves. The electron in hydrogen is confined by the electric force of the nucleus, in what physicist Jay Orear called a "fuzzy walled box". Even though the walls are not rigid, the standing waves have precise shapes.

Although the standing wave patterns we will discuss were calculated for the hydrogen atom, the general features of these patterns apply to the electrons in largeratoms. We will find that when we include Pauli's exclusion principle and the concept of electron spin, we can begin to see how the electron wave patterns determine the chemical properties of atoms and the structure of the periodic table.

## SOLUTIONS OF SCHRÖDINGER'S EQUATION FOR HYDROGEN

In the Bohr theory, the energy levels for hydrogen were determined by the assumption that the electron's angular momentum L was quantized in units of  $\hbar$ . The electron's angular momentum was  $L_1 = \hbar$  in the lowest energy level,  $L_2 = 2\hbar$  in the second level, etc. Assuming circular orbits, and applying classical mechanics, this led to the set of energy levels  $E_n$  given by

$$E_{1} = \frac{-e^{4}m}{2\hbar^{2}} = -13.6 \text{ eV}$$
$$E_{n} = \frac{E_{1}}{n^{2}}$$
(1)

which gave us the values  $E_1 = -13.6 \text{ eV}$ ,  $E_2 = -3.40 \text{ eV}$ , etc., that explained the hydrogen spectra.

De Broglie's contribution was to show that one could understand the reason for quantization of angular momentum by assuming that the electron had a wave nature, with the electron's wavelength  $\lambda$  related to its momentum p by the formula

$$\lambda = \frac{h}{p} \quad de \, Broglie \, formula \tag{2}$$

The quantization of angular momentum came from the picture that an integral number of wavelengths fit around one of the allowed circular orbits.

Following Debye's suggestion (see the introduction to Chapter 37), Schrödinger developed a wave equation with which he was able to solve for the allowed standing wave patterns of the electron in a hydrogen atom. Doing this required no arbitrary assumptions like circular orbits or the quantization of angular momentum. The wave patterns are simply solutions of the wave equation.

Schrödinger's equation has a surprisingly large number of solutions for the allowed standing waves of an electron in hydrogen. These waves are characterized by three numbers commonly given the names "n", " $\ell$ ", and "m". It turns out that for the wave to be an acceptable solution, a solution that does not have an infinite value at some point, the numbers n,  $\ell$ , and m have to have integer values. These integer numbers have become known as *quantum numbers*. Each of the allowed standing wave patterns in hydrogen has a distinct set of values of the quantum numbers n,  $\ell$ , and m. Figure (1) shows six of the allowed patterns. What we have drawn is the intensity of the wave pattern as it would be seen if we looked through the wave. When the side and top view are different we show both to help visualize the three dimensional structure of the wave.

The pattern on the bottom row labeled by the quantum numbers (n = 1, l = 0, m = 0) is a spherical ball with a fuzzy edge. The radius of the ball is about equal to the Bohr radius of .529 angstroms. Schrödinger's equation allows us to calculate the energy of the electron in this pattern and the result is -13.6 eV, the same as the lowest energy state of the electron in the Bohr theory. This is the standing wave pattern for an electron in the ground state – cool, transparent hydrogen.

On the second row in Figure (1) there are four distinct patterns, all with n = 2 but with different values of  $\ell$  and m. Schrödinger's equation predicts that the energy of an electron in hydrogen is given, in general, by the formula

$$E_n = \frac{E_1}{n^2}; \qquad E_1 = \frac{-e^4m}{2\hbar^2}$$
 (3)

where n is the "n" quantum number we have been discussing. Since these are the same values we got from the Bohr theory, Schrödinger's equation predicts all the energy levels needed to explain the entire spectrum of light radiated by hydrogen. Because of Equation 3, it is reasonable to call n the *energy quantum number* for hydrogen.

The first big surprise from Schrödinger's equation is that we can have several standing wave patterns all representing an electron with the same energy. In the n = 2 energy level, there are four distinct patterns representing an electron with the energy  $E_2 = -3.40$  eV. One of these patterns has quantum numbers l = 0, m = 0. The other three have quantum numbers l = 1, m = 1, m = 0, and m = -1.

When we get up to the third energy level, n = 3, there are nine patterns all with an energy of -1.51 eV. There is one pattern with  $\ell = 0$ , m = 0; three patterns with  $\ell = 1$ , m =(1, 0, -1); and five patterns with  $\ell = 2$ , m = (2, 1, 0, -1, -2). As we go up in energy, we get an ever increasing number of patterns. The general rule is that  $\ell$  can range from zero up to n - 1, and the m values can range from  $+\ell$  to  $-\ell$ .



There are 8 more n = 3 patterns in addition to the one shown. The  $\ell$  and m quantum numbers are  $\ell = 1$ ; m = 1, 0, -1 $\ell = 2$ ; m = 2, 1, 0, -1, -2.

 $n = 3, \ \ell = 0, \ m = 0$ 



E = -3.40eV



 $n = 1, \ \ell = 0, \ m = 0$ 

Figure 1

The lowest energy standing wave patterns in hydrogen. The intensity is what you would see looking through the wave.

## **Exercise 1**

a) For the n = 4 energy level, where  $E_4 = -.85eV$ , there are 16 allowed standing wave patterns. What are the values of the quantum numbers  $\ell$  and m for these patterns?

b) How many allowed patterns are there, what are the values of  $\ell$  and m, and what is the energy, for the n = 5 standing wave patterns?



#### Figure 2

Comparison of the L = 0 electron standing wave patterns with the guitar standing waves. Each step up in energy level or harmonic introduces one more node.

## The $\ell = 0$ Patterns

For each energy level n, there is one wave pattern with  $\ell = 0$ . We have shown the first three  $\ell = 0$  patterns in Figure (1). All  $\ell = 0$  wave patterns are spherically symmetric. The n = 1,  $\ell = 0$  pattern, for the ground state electron, is a fuzzy spherical ball with a diameter of about one angstrom. The n = 2,  $\ell = 0$  pattern is a spherical ball surrounded by a spherical shell. Between the ball and the shell, at a radius r = 1.06 angstroms, the wave has value of zero. We can call this a spherical node.

In the n = 3, l = 0 pattern we have a spherical ball surrounded by two spherical shells. There are now two spherical nodes, the inner one located at r = 1.00angstroms and the outer one at r = 3.75 angstroms. As we go up higher in energy, we get one more spherical node for each step up in energy level. In Figure (2), we compare the three lowest energy l = 0 wave patterns with the three lowest frequency standing wave patterns on a guitar string. While the patterns look quite different, both have the feature that as we go up one step in energy or frequency, we get one more node in the wave pattern. The first harmonic (n = 1) has no nodes between the ends of the string. The second harmonic (n = 2) has one node, while the third harmonic (n = 3) has two nodes, etc.

One of the key features of angular momentum is that it represents a rotation about an axis. In Newtonian mechanics, we defined the direction of the angular momentum vector  $\vec{L}$  as being the direction of the axis of rotation. Because all the hydrogen wave patterns with  $\ell = 0$  are spherically symmetric, they have no preferred axis about which the electron could have angular momentum.

## The $\ell \neq 0$ Patterns

In all the  $\ell \neq 0$  patterns, like the three n = 1,  $\ell = 1$  patterns shown in the middle row of Figure (1), there is a special axis about which the electron can have angular momentum. This suggests that the  $\ell$  quantum number is related to the electron's angular momentum, and that when  $\ell = 0$ , the electron has no angular momentum. This is different from the Bohr picture where the electron's angular momentum started with one unit  $L = \hbar$  in the lowest energy level, two units  $L = 2\hbar$  in the second level, etc. The Bohr theory did not allow for zero orbital angular momentum orbits while Schrödinger's equation tells us that there is a zero angular momentum wave pattern in each energy level.

## Intensity at the Origin

Another general feature of the hydrogen wave patterns is that all  $\ell = 0$  patterns have a maximum intensity at the origin, at the nucleus, while all the  $\ell \neq 0$  patterns have a node there. The node at the origin for  $\ell \neq 0$  patterns has a simple classical explanation. The classical formula for angular momentum is *the linear momentum p times the lever arm*  $r_{\perp}$ . In order for the electron to have non zero angular momentum about the nucleus, it must have a non zero lever arm  $r_{\perp}$  and therefore cannot be at the nucleus. (One has to be careful applying Newtonian arguments to atomic phenomena. In the next chapter we will see a similar argument fail when we discuss electron spin).



Figures 1c,d,g, and h repeated

Later in this chapter we will see that the fact that l = 0 patterns have a maximum at the nucleus while the  $l \neq 0$  patterns have a node there, plays an important role in the electron structure and chemical properties of atoms.

## **Quantized Projections** of Angular Momentum

A clue to understanding the  $\ell \neq 0$  wave patterns can be obtained from a more detailed look at the two doughnut shaped patterns in Figure (1), the patterns labeled by the quantum numbers n = 2,  $\ell = 1$ , m = +1 and m = -1. While the m = +1 and m = -1 patterns look the same, a more detailed calculation with the Schrödinger equation shows in the m = +1 pattern the electron is traveling around the doughnut in a *counterclockwise direction*, while in the m = -1 pattern the electron is traveling *clockwise*.

These two patterns have an axis of symmetry which we have labeled the *z* axis, that passes up through the center of the doughnut. (These axes are shown as white dotted lines in the side views of these patterns, Figures 1d and 1h.) Further calculation with Schrödinger's equation shows that the electron in the l = 1, m = 1 pattern (counterclockwise motion), the electron has a *z* component of angular momentum precisely equal to one unit h.

$$L_{z} \begin{pmatrix} \text{for the } \ell = 1 \\ m = 1 \text{ pattern} \end{pmatrix} = \hbar$$
(4a)

For the clockwise motion, the l = 1, m=-1 pattern, the z component of angular momentum is minus one unit  $\hbar$ 

$$L_{z} \begin{pmatrix} \text{for the } \ell = 1 \\ m = -1 \text{ pattern} \end{pmatrix} = -\hbar$$
(4b)

The pattern in between, the one that looks like two tennis balls, one on top of the other, described by the quantum numbers  $\ell = 1$ , m = 0 turns out to have no angular momentum in the z direction.

$$L_{z} \begin{pmatrix} \text{for the } \ell = 1 \\ m = 0 \text{ pattern} \end{pmatrix} = 0$$
(4c)

We see that the "m" quantum number tells us how many units of angular momentum the electron has in the z direction. There is somewhat of an analogy between the three n=2, l = 1 patterns in Figure (1) and the bicycle wheel demonstration we discussed in Chapter 7 (Figures 7-15 and 7-16). In Figure (3) we compare the m = 1 pattern with a bicycle wheel whose angular momentum  $\vec{L}$  points in the + z direction, the m = -1 pattern with a bicycle wheel whose angular momentum points in the -z direction, and the m=0 pattern with a bicycle wheel that has no component of angular momentum in the z direction.

The analogy shown in Figure (3) actually demonstrates how different angular momentum is on an atomic scale from what we are familiar with on a human scale. The most striking difference is that you can point a bicycle wheel in any direction you want. By turning the wheel over, you can change the z component  $L_z$  from +L when it is pointing up to any value down to -L when the wheel is pointing down. Any value between +L and -L is allowed.

For the hydrogen atom, an electron in the second energy level has only three  $\ell = 1$  wave patterns, only three distinct z projections of angular momentum, each differing by one unit of angular momentum  $\hbar$ . There is no wave pattern for L<sub>z</sub> equal to some fractional value of  $\hbar$ —*the projections of angular momentum are quantized!* There is absolutely nothing in Newtonian mechanics that prepares us for understanding how projections of angular momentum can be quantized. It is strictly a consequence of the wave nature of the electron, and the fact that a confined wave has only certain allowed standing wave patterns.

## Figure 3

There are three standing wave patterns for a second energy level, unit angular momentum electron. Schrödinger's equation tells us that the z axis projection of angular momentum in the three patterns are 1 unit, 0 units, and -1 units. There are no intermediate values. because there are no other wave patterns. In comparison, a bicycle wheel has not only the three projections of angular momentum shown, but also manv intermediate values.





















## The Angular Momentum Quantum Number

In Figure (3) we showed the different orientations of a bicycle wheel with a total angular momentum of magnitude L. The z component of the bicycle wheel's angular momentum ranges from +L when the axis is pointing up to -L when the axis is pointing down. For the hydrogen electron in Figure (3),  $\ell = 1$  for all the patterns, and the z projection of the electron's angular momentum ranges from +1 unit for the m = 1 pattern down to -1 unit for the m = -1 pattern. This suggests that the quantum number  $\ell$  represents the total angular momentum of the electron while m represents the allowed z projections.

This interpretation is almost right. The  $\ell$  quantum number is related to the electron's total angular momentum, but the value of the total angular momentum is not quite equal to  $\ell$  units of angular momentum as one might expect. Solving Schrödinger's equation for the magnitude L of the electron's orbital angular momentum about the proton gives the result

$$L = \sqrt{\ell(\ell+1)} \hbar \qquad \begin{array}{c} \text{total angular} \\ \text{momentum of} \\ \text{the electron} \end{array} \tag{5}$$

For large values of  $\ell$ , the difference between  $\ell$  and  $\sqrt{\ell(\ell+1)}$  is slight and the z projections of angular momentum can range essentially from  $+\ell\hbar$  to  $-\ell\hbar$  as one would expect from the bicycle wheel analogy. But for small (non zero) values of  $\ell$ , the total angular momentum is significantly larger than the maximum z projection. For  $\ell = 1$ , the maximum z projection is  $\hbar$ , while the total angular momentum is  $L = \sqrt{1(1+1)} \hbar = \sqrt{2} \hbar$ .

There was no guarantee that angular momentum had to behave on an atomic scale, just the way we expected it to from our experience with large scale phenomena. All we need to do is understand the transition from large to small scale phenomena. In the case of angular momentum, we can picture the bicycle wheel as having a huge angular momentum quantum number  $\ell$ . As a result there are a huge number of allowed projections, with m ranging from  $+\ell$  to  $-\ell$ , which allows us to rotate the bicycle wheel axis in an apparently continuous fashion. And there is essentially no difference between  $\ell$  and  $\sqrt{\ell(\ell+1)}$ , thus the maximum z projection of angular momentum essentially equals the total angular momentum L.

## **Other notation**

Further notation that some readers may have encountered, are the names (*s waves*) for the l = 0 patterns, (*p waves*) for l = 1 patterns and (*d waves*) for l = 2 patterns. These names, which are fairly common, have a rather obscure historical origin. Using this notation, one can, for example, refer to an electron in an n = 3, l = 2 pattern as a 3d wave. The ground state of hydrogen is a 1s wave.

## Exercise 2

Using the s,p,d notation, what would we call the waves shown in Figure (3) ?
# An Expanded Energy Level Diagram

In our discussion of the Bohr theory, we drew an energy level diagram so that we could study transitions from one level to another in order to predict the energy of the photons the atom could emit. The diagram, like the one in Figure (35-3), is quite simple with one line for each energy level.

With the Schrödinger equation we discover that there are numerous standing wave patterns for each energy level. The simple energy level diagram of Figure (35-3) does not give a hint of the multiple wave patterns. Only the energy quantum number n is shown, there being no indication of the  $\ell$  and m quantum numbers (which were unknown when Bohr developed his theory).

It is traditional (and convenient) to expand the energy level diagram as we have done in Figure (4), to distinguish not only the energy quantum numbers n, but also the angular momentum quantum numbers  $\ell$ . We might

be tempted to expand the diagram further and display the separate projections m, but this would make the diagram too complex. (In Figure (4) we indicated the z projections by including some sketches of the lower energy wave patterns. Such sketches are not usually included in energy level diagrams.)

One advantage of the expanded energy level diagram is that it illustrates graphically that the maximum value of  $\ell$  goes up only to n–1. It shows that there is one  $\ell = 0$  pattern for n = 1, an  $\ell = 0$  and an  $\ell = 1$  pattern for n = 2, etc. When you look at this diagram, you have to remember that for each line, the z projections m can range from m = + $\ell$  down to m = - $\ell$  in unit steps.

Another advantage of the energy level diagram of Figure (4) is related to the fact that when an electron in an atom radiates a photon, the electron's  $\ell$  value almost always changes by one unit. This is because a photon carries out angular momentum, and to conserve angular momentum, the electron's angular momentum has to change. The common transitions represent not only steps up and down, but one step sideways.

(It is not impossible for an electron to emit a photon and not change its angular momentum  $\ell$ , it just a much less likely event. We only see such  $\Delta \ell = 0$  transitions, called "forbidden transitions", when the electron has no where to jump and change its  $\ell$  value by one unit. For example, if the electron, for some reason, ends up in the n = 2,  $\ell = 0$  state, the only lower energy state is the n = 1,  $\ell = 0$  state. The electron cannot fall there and change  $\ell$  by one unit. As a result the electron hangs up in the n = 2,  $\ell = 0$  state for a much longer time than it would if a  $\Delta \ell = 1$  transition were available.)



#### **MULTI ELECTRON ATOMS**

Straightforward techniques can be used to solve Schrödinger's equation for one electron atoms like hydrogen. To deal with a two electron atom like helium, we have to take into account not only the attraction between the electrons and the nucleus, but also the repulsion between electrons. This makes Schrödinger's equation more difficult to solve. One has to either use approximation techniques or a computer.

However, for all atoms, there are certain properties that can be understood in terms of the general structure of the hydrogen standing wave patterns, rather than from detailed calculations. We can learn enough from these general properties to begin to see why atoms behave as they do in chemical reactions.

To study multi electron atoms, imagine that we start with hydrogen and add electrons one at a time (also increasing the number of protons and neutrons in the nucleus to keep the atom electrically neutral and the nucleus stable). We will assume that as we add each electron, it falls down to the lowest energy wave pattern available.

If we start with a nucleus with one proton, and drop in one electron, the electron eventually falls down to the  $E_1$ ,  $\ell = 0$  standing wave pattern shown in Figure (1a).

Add a proton to form a helium nucleus, drop in another electron, and we can expect the electron to also fall down to the lowest energy  $E_1$  standing wave pattern. The extra Coulomb attractive force of the two protons in the nucleus strengthens the binding of the electrons, but the repulsive force between the two electrons weakens it. Experimentally, it takes 24.6 eV to remove an electron from helium, while only 13.6 eV are needed for hydrogen. Thus the electrons are more tightly bound in helium, and we see that the extra Coulomb attraction to the nucleus is more important than the repulsion between electrons.

Using helium as a guide, we should expect that when we go to lithium with 3 protons in the nucleus, the increased Coulomb attraction to the nucleus should cause lithium's three electrons to be even more tightly bound than helium's two. This would lead us to predict that it takes even more than 24.6 eV to pull one of the electrons out of lithium.

This is not a good prediction. Experimentally, the amounts of energy needed to remove electrons from lithium one at a time are 5.39 eV, 75.26 eV and 121.8 eV. While two of lithium's electrons are tightly bound, one is very loosely bound, requiring less than half the energy to remove than the hydrogen electron. A possible explanation for the loose binding of lithium's third electron is that, for some reason, that electron did not fall down to the lowest energy  $E_1$  type of standing wave pattern. It appears to be hung up in the much higher energy, less tightly bound  $E_2$  type of standing wave, one of the four  $E_2$  patterns seen in Figure (1).

#### **Pauli Exclusion Principle**

But why couldn't the third lithium electron fall down to the low energy  $E_1$  pattern? In 1925, two separate ideas provided the explanation. Wolfgang Pauli proposed that no two electrons were allowed to be in exactly the same state. This is known as the **Pauli exclusion principle**. But the exclusion principle seems to go too far, because in helium, both electrons are in the same  $E_1$ ,  $\ell = 0$  standing wave pattern. If you cannot have two electrons in exactly the same state in an atom, then something must be different about the two electrons in helium.

#### **Electron Spin**

To explain what the difference between the two electrons might be, two graduate students, Samuel Goudsmit and George Uhlenbeck, proposed that the electron was like a spinning top with its own internal angular momentum. This became known as *spin* angular momentum. The special feature of the electron's spin is that it has *two allowed projections*, which we call *spin up* and *spin down*. In helium you could have two electrons in the same  $E_1$  wave pattern if they had different spin projections, for then they would not be in identical states.

Because the electron spin has only two allowed projections, we cannot add a third electron to the  $E_1$  wave pattern. Lithium's third electron must stop at one of the higher energy  $E_2$  standing wave patterns. Its energy is much less negative and therefore this electron is much less tightly bound than the first two electrons that went down to the  $E_1$  wave pattern.

# THE PERIODIC TABLE

As we go to larger atoms, adding electrons one at a time, the  $E_2$  standing wave patterns begin to fill up. Since there are four  $E_2$  patterns, each with two allowed spin states, up to 8 electrons can fit there. When the  $E_2$ patterns are full, when we get to the element neon with two  $E_1$  electrons and eight  $E_2$  electrons, we have an inert noble gas that is chemically similar to helium. Adding one more electron by going to sodium, the eleventh electron has to go up into the  $E_3$  energy level since both the  $E_1$  and  $E_2$  patterns are full. This eleventh electron in sodium is loosely bound like the third electron in lithium, with the result that both lithium and sodium have similar chemical properties. They are both strongly reactive metals.

Table 1 shows the electron structure and the binding energy of the last electron for the first 36 elements in the periodic table. The general features of this table are that the lowest energy levels fill up first, and there is a large drop in binding energy when we start filling a new energy level. We see these drops when we go from the inert gases helium, neon, and argon to the reactive metals lithium, sodium and potassium. We can see that this sudden change in binding energy leads to a significant change in the chemical properties of the atom.

A closer look at Table 1 shows that there is a relatively steady uniform increase in the electron binding as the energy level fills up. The binding energy goes from 5.39 eV for lithium in fairly equal steps up to 21.56 eV for neon as the  $E_2$  energy level fills. The pattern is more or less repeated as we go from 5.14 eV for sodium up to 15.76 eV for argon while filling the  $E_3$  energy level. It repeats again in going from 4.34 eV for potassium up to the 14.00 eV for krypton.

A closer look also uncovers some exceptions to the rule that the lower energy levels fill first. The most notable exception is at potassium, where the  $E_4$  patterns with  $\ell = 0$  begin to fill before the  $E_3$  patterns with  $\ell = 2$ .

To understand why the binding energy gradually increases as an energy level fill up, and why the  $E_3$ ,  $\ell = 2$  patterns fill up late, we have to take a closer look at the structure of the electron wave patterns and see how this structure affects the binding energy. To do this it is useful to introduce the concepts of electron screening and effective nuclear charge.

#### **Electron Screening**

In our discussion of the binding energy of the two electrons in helium, we pointed out that there was a competition between the increased Coulomb attractive force to the nucleus and the repulsion between the electrons. We could see that the increased attraction was more important because helium's two electrons are each more tightly bound to the nucleus than hydrogen's one. It requires 24.5 eV to remove an electron from helium and only 13.6 eV from hydrogen.

The following argument provides an explanation of this increased binding of helium's electron. Since the two electrons are in the same  $E_1$  wave pattern, half the time a given electron is closer to the nucleus than its partner and feels the full force of the nuclear charge +2e. But half the time it is farther away, and the net charge attracting it toward the nucleus is +2e reduced by the other electron's charge -1e for a total +1e. Thus, on the average the electron sees an effective charge of approximately 1.5 e. This is greater than the charge +1e seen by the single electron in hydrogen, and thus results in a stronger binding energy. What we have done is to account for the repulsion of the other electron by saying that the other electron screens the nucleus, reducing the nuclear charge from +2e to an effective value of approximately 1.5e.

			Binding									
			energy of last	E <sub>1</sub>	E	2		E <sub>3</sub>		E	4	Energy level E <sub>n</sub>
Z		Element	electron in eV	0	0	1	0	1	2	0	1	Angular momentum quantum number {
1 2	H He	Hydrogen Helium	13.60 24.58	1 2								
3 4 5 6 7 8 9 10	Li Be B C N O F Ne	Lithium Beryllium Boron Carbon Nitrogen Oxygen Fluorine Neon	5.39 9.32 8.30 11.26 14.54 13.61 17.42 21.56	Helium core	1 2 2 2 2 2 2 2 2 2 2	1 2 3 4 5 6						
11 12 13 14 15 16 17 18	Na Mg Al Si P S Cl A	Sodium Magnesium Aluminum Silicon Phosphorus Sulfur Chlorine Argon	5.14 7.64 5.98 8.15 10.55 10.36 13.01 15.76	Neo	on core		1 2 2 2 2 2 2 2 2 2 2	1 2 3 4 5 6				
19         20         21         22         23         24         25         26         27         28         29         30         31         32         33         34         35         36	K Ca Sc Ti V Cr Mn Fe Co Ni Cu Zn Ga Ge As Se Br Kr	Potassium Calcium Scandium Titanium Vanadium Chromium Manganese Iron Cobalt Nickel Copper Zinc Gallium Germanium Arsenic Selenium Bromide Krypton	$\begin{array}{c} 4.34\\ 6.11\\ 6.56\\ 6.83\\ 6.74\\ 6.76\\ 7.43\\ 7.90\\ 7.86\\ 7.63\\ 7.72\\ 9.39\\ 6.00\\ 7.88\\ 9.81\\ 9.75\\ 11.84\\ 14.00\end{array}$		Ar	gon c	ore		$ \begin{array}{c} 1\\2\\3\\5\\6\\7\\8\\10\\10\\10\\10\\10\\10\\10\\10\\10\\10\\10\\10\end{array} $	$ \begin{array}{c} 1\\2\\2\\2\\2\\1\\2\\2\\2\\2\\2\\2\\2\\2\\2\\2\\2\\2\\2\\2$	1 2 3 4 5 6	

# Table 1

Electron binding energies. Adapted from Charlotte E Moore, Atomic Energy Levels, Vol II, National Bureau of Standards Circular 467, Washington, D.C., 1952.

# **Effective Nuclear Charge**

To see what effect changing the nuclear charge has on the binding energy, we can go back to our Bohr theory calculations for different one electron atoms. In Exercises 9 and 10 of Chapter 35 we found that the ground state energy of a single electron in an atom where the nucleus had *z protons* was

$$E_1 = z^2 \times (-13.6 \text{ eV}) \qquad \begin{array}{c} \text{ground state} \\ \text{energy in a single} \\ \text{electron atom} \end{array} \tag{6}$$

While Equation 6 was derived from the Bohr theory, it gave results in excellent agreement with experiment as one can easily see from working Exercise 3.

#### **Exercise 3**

Table 2 lists the binding energy for the last electron for the elements hydrogen through boron. This is the binding energy when all the other electrons have already been removed. For each element, check the prediction that the binding energy is given by Equation 6.

		Binding energy			
Ζ	Element	of last electron			
1	Hydrogen	13.6 eV			
2	Helium	54.14 eV			
3	Lithium	121.8 eV			
4	Beryllium	216.6 eV			
5	Boron	338.5 eV			
Table 2					

Equation 6 suggests that if an electron in a multi electron atom sees an effective nuclear charge  $(z_{eff} e)$ , the electron binding energy should be approximately  $z_{eff}^2$ times the energy the electron would have in the same energy level in hydrogen. Trying out this idea on helium, where we estimated  $z_{eff}$  to be about 1.5e, we get

$$E_{1} \begin{pmatrix} neutral \\ helium \end{pmatrix} = z_{eff}^{2} \times (-13.6 \text{ eV})$$
$$= (1.5)^{2} \times (-13.6 \text{ eV})$$
$$= -30.6 \text{ eV}$$
$$= \underset{of helium}{estimated}$$
(7)

This estimate of 30.6 eV is about 25% too high since the experimental value is only 24.6 eV. We can take this to imply that our estimate of  $z_{eff} = 1.5$  e for helium was a bit too crude. Our simple arguments about screening are not a substitute for an accurate calculation using Schrödinger's equation.

What we can do, however, is to turn our approach around and use the experimental values of the binding energy to calculate an effective nuclear charge  $(z_{eff})$ . Doing this for helium gives

$$E_{1} \left( \substack{neutral \\ helium} \right) = z_{eff}^{2} \times \left( -13.6 \text{ eV} \right)$$
$$-24.6 \text{ eV} = z_{eff}^{2} \times \left( -13.6 \text{ eV} \right)$$
$$z_{eff} = \sqrt{\frac{24.6}{13.6}} = 1.34$$
(8)

The value of 1.34 is not too far off our original guess of 1.5. The result tells us that the electron screening is a bit more effective than we had predicted.

#### Lithium

We will now see that various features of the periodic table begin to make sense when viewed in terms of electron screening and the structure of the electron wave patterns. Let us start off with lithium where the last electron is in the  $E_2$ ,  $\ell = 0$  pattern and has a binding energy of 5.39 eV. Since this electron is in an  $E_2$  energy level, our formula for  $z_{eff}$  should be

$$E_{1}(lithium) = z_{eff}^{2} \times (-3.40 \text{ eV})$$
  
-5.39 eV =  $z_{eff}^{2} \times (-3.40 \text{ eV})$   
 $z_{eff} = \sqrt{\frac{5.39}{3.40}} = 1.26$  (9)

where we used - 3.40 eV rather than - 13.6 eV because we are discussing an  $E_2$  electron. In a nucleus with 3 protons, why does the  $E_2$  electron only see an effective charge of 1.26e? The answer lies in the shape of the  $E_1$  and  $E_2$  wave patterns. The first two electrons in lithium are in the  $E_1$  pattern of Figure (1a) reproduced below. It consists of a small spherical ball centered on the nucleus. The third electron, the one whose binding energy we are discussing, is in the  $E_2$ ,  $\ell = 0$  pattern of Figure (1b). This pattern consists of a larger spherical ball surrounded by a spherical shell. The electron in this pattern spends a considerable amount of the time outside the smaller spherical ball of the two  $E_1$  electrons. Thus much of the time the third electron sees only an effective nuclear charge of about (1.0e). Some of the time, however, the third electron is also down at the nucleus feeling the full nuclear charge of (3e). That the average nuclear charge seen by the third electron is (1.26e) is not too difficult to believe.







#### **Beryllium**

When we went from one  $E_1$  electron in hydrogen to two  $E_1$  electrons in helium, the binding energy about doubled, from 13.6 eV to 24.6 eV. In going from one  $E_2$  electron in lithium to two  $E_2$  electrons in beryllium, the binding energy increases from 5.39 eV to 9.32 eV. Again the electron binding energy almost doubled as we went from one to two electrons in the same energy level.

#### Boron

When we go from beryllium to boron, we add a third electron to the  $E_2$  energy level. From our experience with beryllium, we expect another significant increase in binding energy, up to perhaps 13 eV or 14 eV. But instead the binding energy drops from 9.32 eV down to 8.30 eV. Something broke the pattern and caused this drop.

At boron, both the  $E_2$ ,  $\ell = 0$  wave patterns are already full and the electron has to go into one of the  $E_2$ ,  $\ell = 1$ patterns. All the electron standing wave patterns with a non zero amount of angular momentum have a node at the origin. The more the angular momentum, the more spread out the node and the farther the electron is kept away from the nucleus. An electron in an  $\ell \neq 0$  wave pattern will thus be effectively screened by electrons in  $\ell = 0$  wave patterns where the electron spends a lot of time right down at the nucleus. Thus we expect that electrons in  $\ell \neq 0$  patterns to be less tightly bound than those in the  $\ell = 0$  pattern of the same energy level. This shows up with the drop in binding energy in going from beryllium to boron.

# Up to Neon

For all atoms beyond helium, there is a core consisting of the nucleus and the two tightly bound  $E_1$  electrons. As the charge on the nucleus increases, the size of the  $E_1$  patterns shrink, and are penetrated less and less by the outer electrons. We can think of this helium core as acting as the effective nucleus for the larger atoms.

As we go from boron to neon, the charge on the helium core increases from 3e to 8e as the  $E_2$ ,  $\ell = 1$  patterns fill up. This increase in the charge of the core causes a more or less gradual increase in the binding energy, from 8.30 eV up to 21.56 eV. The one exception is the slight drop in binding energy as we go from nitrogen to oxygen. The arguments we have made so far are not detailed enough to explain this drop.

#### Sodium to Argon

We get the expected large drop in binding energy as we go from neon to sodium and start filling the  $E_3$  patterns. The  $E_3$ ,  $\ell = 0$  patterns are full at magnesium and we get a small drop in binding energy as the non zero angular momentum patterns  $E_3$ ,  $\ell = 1$  start to fill at aluminum. Again the angular momentum keeps the electrons away from the nucleus and increases the screening. As the  $E_3$ ,  $\ell = 1$  patterns fill up, they are building a structure on the ever shrinking neon core. The charge on the neon core increases from (3e) at aluminum to (8e) at argon, again causing a gradual buildup of the electron binding energy from 5.98 eV to 15.76 eV. There is even the slight glitch going from phosphorous to sulfur that mirrors the glitch from nitrogen to oxygen.

# **Potassium to Krypton**

The first major break in the pattern of filling the lower energy levels first occurs at potassium. At potassium the  $E_3$ ,  $\ell = 2$  levels remain unfilled while the last electron goes into the higher energy level  $E_4$ ,  $\ell = 0$ pattern. At this point the screening due to angular momentum has become more important than the energy level. The  $\ell = 2$  patterns have such a big fat node at the nucleus that an  $\ell = 2$  electron cannot get near the nucleus to feel the now large nuclear charge. Even though an  $E_4$ ,  $\ell = 0$  electron is in a higher energy level, its wave pattern has a non zero value right down at the nucleus. Some of the time this electron feels the full charge of (19e) for potassium, and this increases the binding beyond that of the  $E_3$ ,  $\ell = 2$  patterns.

At calcium, the  $E_4$ ,  $\ell = 0$  pattern is full, and now the five  $E_3$ ,  $\ell = 2$ ,  $m_{\ell} = +2, +1, 0, -1$ . -2 patterns begin to fill up. There is room for 10 electrons in these 5 patterns, and that takes us down to zinc. As the  $E_3$ ,  $\ell = 2$  patterns fill up underneath the  $E_4$ ,  $\ell = 0$  pattern, there is little change in the outer electron structure and the binding energy increases slowly. The result is that the 10 elements from scandium to zinc have similar chemical properties—all are metals. In some periodic tables, these elements are shown as the first set of *transition elements*.

As we go from gallium to krypton we have the familiar pattern of the  $E_4$ ,  $\ell = 1$  states filling up. There is a gradual increase in binding energy from 6.00 eV at gallium to 14.00 eV at the noble gas krypton. There is even the slight glitch in binding energy going from arsenic to selenium that mirrors the glitches from phosphorus to sulfur, and from nitrogen to oxygen.

#### Summary

At this point it should be clear that the structure of the periodic table of the elements arises from the allowed electron standing wave patterns. Because of the exclusion principle, no two electrons can be in the same state. But because electron spin has two allowed states, up to two electrons can fit into each standing wave pattern.

In general, as we go to atoms with more electrons, the lowest energy patterns fill up first, and there is a significant change in chemical properties when a new energy level begins to fill. But the angular momentum of the wave pattern also plays a significant role. The  $\ell = 0$  patterns can penetrate down to the nucleus, where the electron feels the full strength of the nuclear charge. The  $\ell \neq 0$  patterns have a node at the nucleus, and the full nuclear charge is screened by  $\ell = 0$  electrons.

The effect of angular momentum shows up most noticeably at potassium and calcium, where the two  $E_4$ ,  $\ell = 0$  patterns fill before the  $E_3$ ,  $\ell = 2$  patterns. Because of the extra angular momentum of the  $\ell = 2$  electrons, the  $\ell = 2$  patterns have an extra large node at the nucleus, keeping these electrons farther away and more effectively screened.

As we get to the heavier elements in the periodic table, those beyond krypton, the energy levels get closer together and the binding energy depends more on the detailed structure of the wave patterns. As a result it becomes more difficult to predict how the wave patterns will be filled and to estimate what the binding energies should be. But despite this, we have been able to go a long way in explaining the structure of the periodic table from a few simple arguments about the shape of the electron standing waves in hydrogen, and the idea of electron screening.

# **IONIC BONDING**

In 1871 the Russian chemist Dimitri Mendeleyev worked out the periodic table of the elements from an analysis of the atomic weights and chemical reactions of the elements. Here we will reverse Mendeleyev's approach and use Table 1, our shortened version of the periodic table, to explain some of the typical chemical reactions and chemical compounds.

As an example, suppose we placed a sodium atom next to a chlorine atom, what would happen? The sodium atom has one loosely bound electron in the  $E_2$ ,  $\ell = 0$ wave pattern. The binding energy of this electron is 5.14 eV. The chlorine atom has seven  $E_2$  electrons all tightly bound because of the increase in the effective nuclear charge seen by these electrons. It requires 13.01 eV to remove an electron from chlorine.

If the sodium and the chlorine atom are brought close enough together, the loosely bound outer sodium electron can lose energy by moving into the remaining  $E_2$ wave pattern in the chlorine atom. We end up with a negative chlorine *ion* Cl<sup>-</sup>, where all the  $E_2$  patterns are full, and a positively charged sodium ion Na<sup>+</sup> which has lost its outer electron. These charged ions then attract each other electrically to form a sodium chloride molecule NaC1 which is common table salt.

Sodium chloride is a typical example of *ionic bonding*. The class of elements like lithium, sodium, magnesium, aluminum, etc. that have one, two, or even three loosely bound electrons, tend to give up these electrons in a chemical reaction. These are called *metals*. Those elements like oxygen, fluorine and chlorine, which have nearly full wave patterns and tightly bound electrons, tend to take up electrons in a chemical reaction and are called *non metals*. When metals and non metals combine, held together by ionic bonding, you get a compound called a *salt*.

By looking at the number of loosely bound electrons in a metal, or the number of empty slots in a non metal (the number of electrons required to get to the next noble gas), you can predict the kind of compounds an element can form. For example, sodium, magnesium, and aluminum have one, two and three loosely bound electrons respectively, while oxygen has two empty slots. (Oxygen has six  $E_2$  electrons, and needs two more to fill up the  $E_2$  standing wave patterns). When you completely burn the three metals, the oxides you end up with are  $Na_2O$ , MgO and  $Al_2O_3$ . In  $Na_2O$ , two sodium atoms each contribute one electron to fill oxygen's two slots. In MgO magnesium's two loosely bound electrons are taken up by one oxygen atom. In Al<sub>2</sub>O<sub>3</sub>, two aluminum atoms each supply three electrons, these six electrons are then taken up by three oxygen atoms. There is no simpler way for all the aluminum's loosely bound electrons to completely fill all of oxygen's empty slots.

Hydrogen has one moderately bound electron which it can give up in some chemical reaction and act like a metal. An example is hydrochloric acid, HCl, where the chlorine ion has grabbed the hydrogen electron.

Hydrogen can also behave as a non metal. When it combines with active metals like lithium and sodium, hydrogen grabs the metal's loosely bound electron to complete its  $E_1$  standing wave pattern. The results are the compounds lithium and sodium hydride, LiH and NaH.

More important to life are the bonds like those between hydrogen and carbon atoms which are not ionic in nature. Neither atom has a strong preference to give up or grab electrons. Instead the bonding results from the sharing of electrons. This is the *covalent bonding* that we described in our discussion of the hydrogen molecule in Chapter 18.

# Chapter 39 Spin

In the last chapter we saw that the basic structure of the periodic table follows from the idea that up to two electrons can fit into any given standing electron wave pattern. If you consider the Pauli exclusion principle which says that two electrons cannot be in exactly the same state, then you have to find some difference between the two electrons that can occupy the same wave pattern. Gaudsmit and Uhlenbeck introduced the concept of electron spin to explain this difference. They proposed that the electron had an inherent angular momentum or spin that had two allowed projections, and that the difference between the two electrons in one wave pattern was their spin projections. We commonly call these two allowed projections **spin up** and **spin down**.

In this chapter we take a more detailed look at electron and nuclear spin and the interaction of spin with a magnetic field. An electron with its spin projection parallel to the magnetic field gains magnetic energy, while the opposite projection loses it. The amount of energy gained or lost is proportional to the strength of the magnetic field.

The most accurate way to measure the spin magnetic energy, is to start with an electron in the low energy state and strike it with a photon. If the energy of the photon is precisely equal to the energy required to raise the electron from the low energy spin projection to the high energy spin projection, the photon can be absorbed. We say that the photon flips the spin of the electron. Since the energy of a photon is proportional to its frequency according to Einstein's photoelectric effect formula E=hf, measuring the frequency of the electromagnetic radiation that causes a spin flip tells you how big the magnetic energy is.

The energy required to flip the spin of an electron is usually not very large. If the electron is in a magnetic field of around 10 gauss, typical fields produced by the Helmholtz coils used in several of our laboratory experiments, then photons in radio waves whose frequency is of the order of 30 megacycles have enough energy to flip the electron spin. Since it is not hard to generate electromagnetic waves of this frequency, we can observe electron spin flip using much of our standard laboratory equipment.

When we talk of electromagnetic waves with frequencies of the order of 30 megacycles, we are talking about radio waves between the AM and FM broadcast bands. It is such a low frequency that individual photons should be very hard to detect. Yet the spin flip experiment does just that.

At radio wave frequencies, Maxwell's theory and the ideas of classical electric and magnetic fields should work as well as the photon picture. If we treat the spinning electron as a classical gyroscope with a magnetic moment, we find that a magnetic field can exert a torque on the gyroscope, causing the gyroscope to precess. If we add an oscillating magnetic field, oscillating at the frequency of precession, essentially pushing on the gyroscope once each time it comes around, the gyroscope can gain energy from the oscillating field. This is a resonance phenomena; it is like pushing a kid on a swing. You have to push the kid in time with the swing in order to increase the amplitude of the motion.

It turns out that the frequency with which we have to oscillate the magnetic field is the same as the frequency of the photon that can cause the electron spin to flip. The quantum picture of a photon flipping a spin, and the classical picture of a precessing gyroscope in resonance with an oscillating magnetic field, gives the same results. Thus it is a matter of convenience whether you use the classical or quantum picture.

Because the classical picture involves a resonance, this spinflip process is called **electron spin resonance**. In this chapter we discuss an electron spin resonance experiment.

For a given magnetic field, much less energy is required to flip the spin of a nucleus than of an electron. Measurements of nuclear spin flip energies, in the socalled **nuclear magnetic resonance** experiments, can be done so accurately that one can study not only the spin of the nucleus but also the magnetic environment in which the nucleus sits. Nuclear magnetic resonance forms the basis of magnetic resonance imaging which has become such an important diagnostic tool in medicine.

# THE CONCEPT OF SPIN

A spinning top has an inherent angular momentum, but if you try to picture an electron as a spinning top, you run into conceptual problems. First of all, if you have a spinning top, you can orient the top in any direction you please. The top's angular momentum vector can point up, down, sideways to the left, sideways to the right. But when we describe the electron's spin, we have only two orientations, up and down. We ran into the puzzling idea of quantized projections of angular momentum in our interpretation of the allowed standing wave patterns of the hydrogen atom. But the idea of the electron's spin or rotational axis only pointing up or down seems even more counter intuitive.

Another blow at our classical intuition for angular momentum is our current theoretical picture of the electron as a point particle. No experiment has demonstrated any finite size to the electron, and the theory that treats the electron as a point particle, quantum electro-dynamics, is the most accurately tested theory in all of physics. (String theory allows for some size for an electron, radii of the order of  $10^{-72}$  cm, but there are no experimental tests of string theory.)

If an electron has no radius, how can it have an inherent angular momentum? Angular momentum is linear momentum times a *lever arm*. How can there be angular momentum if the particle has no radius, no lever arm? The classical picture of electron spin resembling that of a spinning tops leaves a lot to be desired. However, despite the problems one encounters, this picture does lead to some useful insights which we will mention shortly. Perhaps the best way to view electron spin is to realize that we are dealing with a wave equation, and wave equations have specific allowed standing waves as solutions. While electron spin does not come from Schrödinger's wave equation, it does from Dirac's more accurate relativistic wave equation. From Dirac's equation, we find that the electron has an inherent angular momentum of  $\hbar/2$ , with two possible projections along the z axis,  $+\hbar/2$  and  $-\hbar/2$ . These are the two allowed states of the electron. They are not different standing wave shapes like the hydrogen standing waves of Figure (1) of the last chapter, but they are different solutions to Dirac's wave equation.

One of the surprises is that the spin angular momentum of the electron is half a unit  $\hbar/2$ . The quantity  $\hbar$  is not the smallest amount of quantized angular momentum,  $\hbar/2$  is. The standard terminology is to say that the electron has half a unit of angular momentum, that it is a "spin 1/2" particle. The orbital angular momentum, representing the motion of the electron around a nucleus, is quantized in units of  $\hbar$ . Only spin angular momentum can come in half integer units.

While the electron's spin has a half integer value, its projection along the z axis changes by an integer value. In our discussion of the angular momentum of the hydrogen standing waves, we saw that an electron in a certain energy level  $E_n$ , with a total angular momentum  $\ell$ , could have z projections ranging from  $\ell$ ,  $\ell - 1$ ,  $\ell - 2$  down to  $-\ell$ . The allowed projections changed in units. The same is true for the electron spin, the allowed projections are +1/2 and -1/2, a change of one unit.

# INTERACTION OF THE MAGNETIC FIELD WITH SPIN

One of the predictions of the Dirac equation for electrons is that the electron spin interacts with a magnetic field. The state with the spin parallel to the magnetic field gains magnetic energy while the other state loses it. The amount of energy gained or lost is proportional to the strength B of the magnetic field, and the proportionality constant turns out to be a quantity called the *Bohr magneton*, designated by the symbol  $\mu_B$ .

$$\begin{array}{l} \begin{array}{l} \begin{array}{l} \begin{array}{l} magnetic\\ energy \ of\\ electron\\ spin \end{array} & E_{mag} \end{array} = \begin{cases} +\mu_{B}B \ \text{spin parallel to } \vec{B}\\ -\mu_{B}B \ \text{spin opposite to } \vec{B} \end{array}$$
(1)

$$\mu_{\rm B} = e\hbar/2m$$
 Bohr magneton  
= 5.79 × 10<sup>-5</sup> eV/tesla

The amount of energy required to flip an electron from its low energy state to the high energy state is thus

$$\Delta E_{mag} = 2\mu_B B$$
energy required to
flip the electron spin
in a magnetic field
(3)

(2)

Since a Bohr magneton is  $5.79 \times 10^{-5}$  eV/tesla we can express Equation 3 numerically as

$$\Delta E_{\text{mag}} = \left(11.6 \times 10^{-5} \text{B}\right) \text{eV}$$
(4)

where B has to be expressed in tesla.

If you wish to measure magnetic fields in gauss, then convert  $\mu_B$  to eV/gauss:

$$\mu_{\rm B} = 5.79 \times 10^{-5} \frac{\rm eV}{\rm tesla} \times \frac{1}{10^4 \frac{\rm gauss}{\rm tesla}}$$
$$\mu_{\rm B} = 5.79 \times 10^{-9} \frac{\rm eV}{\rm gauss} \tag{5}$$

# Magnetic Moments and the Bohr Magneton

The formula for the Bohr magneton has its origin in a combination of classical physics with the Bohr theory. Back in Chapter 31 we observed that if you place a loop of wire in a magnetic field  $\vec{B}$ , and then run an electric current i through the wire, the magnetic field can exert a torque on the loop. We found that if you curled the fingers of your right hand in the direction the current is going around the loop, then the magnetic torque tended to orient the loop so that your thumb pointed parallel to the magnetic field. We called this the low energy orientation of the loop. To turn the loop over to the high energy orientation required an amount of work that was proportional to the current i, the strength of the magnetic field B, and to the area A of the loop. The explicit formula for the amount of work required was

energy required  
to turn loop over 
$$= 2(iA)B$$

We defined the product of the current i times the area A as the *magnetic moment*  $\mu$  of the current loop.

$$\mu \equiv iA \qquad (see 31-34)$$

which gave us the formula

$$\begin{array}{c} \text{energy required} \\ \text{to turn loop over} \end{array} = 2 \mu B \qquad (31-36a)$$

Later in the chapter we considered a special kind of current loop consisting of a charge q moving at a speed v in a circular orbit of radius r. We found that the magnetic moment  $\mu = iA$  of this special current loop could be written in the form

$$\mu = \frac{q}{2m} (mvr) \tag{31-38a}$$

However mvr is the angular momentum L (we called it J back there because we were using L for inductance). Thus the formula for the magnetic moment  $\mu$ of an orbiting charge can be written

$$\mu = \frac{q}{2m}L\tag{31-39}$$

The above result is strictly classical. If we jump ahead to the Bohr picture where we are dealing with an electron whose charge is q = -e, and angular momentum L is quantized in units of ħ, then we find that the magnetic moment is quantized in units of

$$\mu_{\rm B} = \frac{e}{2m}\hbar \tag{6}$$

where this unit of magnetic moment  $\mu_B$  is called a *Bohr* magneton. This is where the name and formula for the Bohr magneton originated. The same constant appeared in Dirac's equation for the energy required to flip the spin of the electron.

The minus sign of the electron charge means that the high energy orientation of the electron is when the spin is parallel to the magnetic field.

#### **Exercise 1**

The formulas

$$E_{magnetic} = 2\mu_B B$$
  
 $\mu_B = \frac{e}{2m}\hbar$ 

came mostly from Chapter 31 where we were working in MKS units. As a result, we need to use MKS units to evaluate  $\mu_B$ . (The constant  $\mu_B$  has a different formula in CGS units.)

We can get the dimensions of  $\mu_B$  from the equation  $E_{magnetic} = 2\mu_B B$ , or

$$\mu_{B} = \frac{E_{magnetic}}{2B} \frac{joules}{tesla}$$

Thus when you use MKS units to evaluate  $\mu_B=e\hbar/2m$ , your answer comes out in joules/tesla rather than eV/tesla. To get the final answer in eV/tesla, you then use the conversion factor  $1.6\times 10^{-19}$  joules/eV.

With this background, show that

$$\mu_{B} = 5.79 \times 10^{-5} \, \frac{\text{eV}}{\text{tesla}}$$

(You will get a value of  $\mu_B = 5.82 \times 10^{-5} \, eV/tesla$ , which differs slightly due to the way we have rounded off the constants.)

# **Electron Spin Resonance Experiment**

The basic idea of the electron spin resonance experiment is to flip the spin of an electron by striking the electron with a photon. The electron's spin will flip only if the photon's energy hf is equal to the magnetic spin flip energy  $2\mu_B B$ . Thus we wish to test the relationship

$$hf = 2\mu_B B \quad spin \ flip \ requirement$$
(7)

where f is the frequency of the photon.

#### **Exercise 2**

a) An electron is placed in a 10 gauss magnetic field. How much energy, in eV, is required to flip the electron from its low energy to its high energy state? (Answer:  $1.15 \times 10^{-7}$  eV).

b) You wish to flip the spin of the electron in part (a) by striking it with a photon. Assume that the photon is absorbed by the electron, and that all the photon's energy goes into flipping the electron's spin. What wavelength photon should you use? (Answer:  $\lambda = 1071 \text{ cm}$ )

c) What is the frequency of the photon in part (b)? (Answer: 28 megacycles).

#### **Exercise 3**

The student FM radio station at Dartmouth College broadcasts on a frequency of 99.4 megacycles. If you wished to use this frequency radiation to flip the spin of an electron in a magnetic field B, what should be the strength of B? Give the answer in gauss.

In our discussion of the particle nature of light, we pointed out that because a radio wave consists of so many photons of such low energy, it would be difficult to detect individual photons, and thus the wave nature of radio waves should predominate. However the spin of the electron is just the right detector for these low energy photons. An electron spin flip experiment can be viewed as an experimental detection of the individual photons in a radio wave.

#### **Nuclear Magnetic Moments**

Both the proton and neutron are spin 1/2 particles, which means that they each have a spin angular momentum with two allowed spin states, spin up and spin down. If you place either of these particles in a magnetic field, one of the projections will gain magnetic energy while the other loses it.

For an electron, the high magnetic energy state was when the spin pointed parallel to the magnetic field. Because the proton has the opposite charge from the electron, the opposite orientation is the high magnetic energy state.

If the Dirac equation is applied accurately to a proton, then the formula for the proton's magnetic moment would be one *nuclear magneton*  $\mu_N$  defined by the equation

$$\mu_{\rm N} = \frac{e\hbar}{2m_{\rm proton}} \qquad \begin{array}{l} definition \ of \ the \\ nuclear \ magneton \end{array} \tag{8}$$

which is the Bohr magneton formula with the electron mass replaced by the proton mass. Since the proton is 1836 times heavier than an electron, a nuclear magneton is 1/1836 times smaller than a Bohr magneton.

The Dirac equation, however, does not give the correct value for the proton's magnetic moment  $\mu_p$ . The experimental value is

$$\mu_{\rm p} = 2.79\,\mu_{\rm N} \tag{9}$$

The fact that the Dirac equation is off by a factor of 2.79 is one indication that the proton is a more complex object than the electron.

(The Dirac equation is not exact even for the electron. The experimental value for the electron's magnetic moment is 1.00114 Bohr magnetons. The correction of .00114 Bohr magnetons is accurately explained by the theory of quantum electrodynamics.)

# Sign Conventions

To handle the fact that the electrons and protons have opposite charges and therefore opposite magnetic moments, the following sign conventions are generally used

$$\begin{array}{l} \underset{\text{energy of spin}}{\overset{\text{magnetic}}{\text{magnetic}}} & E_{\text{magnetic}} = -\vec{\mu} \cdot \vec{B} \\ \\ = \begin{cases} -\mu B & \text{spin parallel to B} \\ +\mu B & \text{spin opposite to B} \end{cases}$$
(10)

where the electron, proton and neutron have the following magnetic moments

$$\mu_{e} = -1.00114 \mu_{B} \quad \substack{electron \\ magnetic moment}$$
(11)

$$\mu_{\rm p} = 2.79\,\mu_{\rm N} \qquad \begin{array}{c} proton\\ magnetic\ moment \end{array} \tag{9}$$

$$\mu_{n} = -1.19 \mu_{N} \qquad \stackrel{neutron}{magnetic moment} \qquad (12)$$

where the Bohr magneton  $\mu_B$  is

$$\mu_{\rm B} = \frac{e\hbar}{2m_{\rm electron}} = 5.79 \times 10^{-5} \frac{eV}{\text{tesla}} \quad (13a)$$

and the nuclear magneton  $\mu_N$  is

$$\mu_{\rm N} = \frac{e\hbar}{2m_{\rm proton}} = 3.15 \times 10^{-8} \frac{eV}{\text{tesla}} \qquad (13b)$$

Note that by putting a - (minus) sign in the formula for  $E_{magnetic}$ , and making the electron's magnetic moment negative, we still have the result that the electron's spin magnetic energy is positive when the electron's spin is parallel to  $\vec{B}$ .

Where the Dirac equation completely fails is in the case of the neutron. If the neutron were a simple uncharged particle, it would have no magnetic moment. The fact that it does have a magnetic moment suggests that, while it has no net charge, it must be some kind of composite object with charged particles inside. We now know that this suggestion is correct. The neutron is made up of three quarks, one *up quark* with a charge +2/3 e and two *down quarks* with a charge -1/3 e. While there is no net charge, the quarks contribute to magnetic energy. (The proton, which consists of two up quarks and one down quark, has a total charge of -2/3e + 2/3e - 1/3e = +e.)

As we mentioned, the difference between the Bohr magneton  $\mu_B$  and the nuclear magneton  $\mu_N$  is due to the mass difference between the electron and the proton. Since a proton is 1836 times as massive as an electron, the nuclear magneton is 1836 times smaller than the Bohr magneton. The result is that the magnetic moments of protons, neutrons, and nuclei in general are typically an order of a thousand times smaller than the electron magnetic moment. To get the same magnetic spin energies as you do for electrons, you thus need magnetic fields of the order of a thousand times stronger when working with nuclei.

#### **Exercise** 4

(a) Express the magnetic moment of the proton in eV/tesla.

(b) A proton is in a 1 tesla magnetic field. How much energy, in eV, is required to flip the spin of the proton?

(c) What is the wavelength and frequency of a photon that can flip the spin of the proton in part (b)? (Answer: (a)  $8.79 \times 10^{-8}$  eV/tesla, (b)  $17.6 \times 10^{-8}$  eV, (c) 706 cm and 42.5 megacycles.)

#### Exercise 5

What strength magnetic field should you use so that photons from the student FM radio station (99.4 megacycles) can flip the spin of the proton? (Answer: 2.34 tesla.)

# **Classical Picture of Magnetic Resonance**

In the appendix to this chapter we work out the classical picture of the interaction of the electron's spin with a magnetic field. One pictures the spinning electron as acting as a tiny current loop with a magnetic moment  $\mu$  as described at the end of Chapter 31. The current loop also has an angular momentum  $\vec{L}$  which makes it act like a gyroscope. If you place the current loop in a magnetic field, as shown in Figure (1), the magnetic field exerts a torque and one predicts that the current loop should precess about the magnetic field lines. The precession is analogous to the precession of the bicycle wheel gyroscope studied in Chapter 12.

In this classical picture, if you subject the precessing current loop to the electromagnetic field of a radio wave, whose frequency f is equal to the precessional frequency  $f_p$  of the loop, the loop can gain energy from the radio wave. This is a resonance phenomena, where the push of the fields of the radio wave have to match the timing of the precession of the loop. It is analogous to pushing a child on a swing, where you have to time your pushes with the motion of the child in order to add energy to the motion.

The classical picture of a precessing current loop gradually gaining energy from a radio wave, and the quantum picture of an electron spin being flipped by a photon, happen to lead to nearly the same predictions. If we start with the condition  $hf = 2\mu_B B$  for the photon energy to match the spin flip energy, then replace  $\mu_B$  by  $e\hbar/2m = e(h/2\pi)/2m = eh/4\pi m$ , we get



In Equation 14, Planck's constant h cancels and we get

$$f = \frac{e}{2\pi m} B \qquad \begin{array}{l} frequency of radio wave \\ photon that can flip \\ an electron spin \end{array} \tag{15}$$

as the relationship between the frequency f of the radio wave and the strength of the magnetic field B.

The fact that Planck's constant cancelled in Equation 15 suggests that a classical analysis might give similar results. In the appendix, we analyze the behavior of a current loop consisting of a particle of charge q and mass m, travelling in a circular orbit. The magnetic moment  $\vec{\mu}$  of the loop points along the axis of the orbit. If the loop is placed in a magnetic field  $\vec{B}$  oriented perpendicular to  $\vec{\mu}$ , then the loop will precess around the magnetic field line at a precessional frequency given by the formula

$$f = \frac{e}{2\pi m} B \qquad precessional frequency of a current loop in a magnetic field (16)$$

which is the same formula as Equation 15 for the frequency of a radio wave photon that can flip the spin of an electron. In the classical picture, if we superimpose a radio wave at this frequency, we get a resonance between the frequency of the radio wave and the precessional frequency of the current loop, enabling the radio wave to add energy to the current loop.

The classical calculations have certain errors that have to be corrected on an ad hoc basis. The current loop model leads to a relationship between the loop's angular momentum L and its magnetic moment  $\mu$ . If we evaluate  $\mu$  experimentally from the relationship  $E_{mag} = -\vec{\mu} \cdot \vec{B}$ , and set  $L = \hbar/2$  for a spin one half particle, the classical relationship is off by a factor of 2 for electrons and 2 x 2.79 for protons. These errors are accounted for by introducing a fudge factor called the *Laudé g factor* to correct the value of  $\mu$ . Since the classical picture has fundamental problems, such as no hint of quantization of angular momentum, and no explanation of how a particle of zero radius can have angular momentum, it is surprising that the semi classical picture works as well as it does.

# ELECTRON SPIN RESONANCE EXPERIMENT

The point of the electron spin resonance experiment is to detect the electron spin flip energy  $\Delta E = 2\mu_B B$ predicted by the Dirac equation using photons of energy E = hf. You might try to do this by placing a container of hydrogen in a magnetic field B and radiating the hydrogen with a radio wave. If the frequency f of the radio wave were such that the photon energies hf equalled the spin flip energy  $2\mu_B B$ , then perhaps we could detect radio wave energy being absorbed as hydrogen atom electrons in the low energy spin state were flipped over to the high energy spin state.

Such an experiment will not work because of an interesting quantum mechanical effect. Hydrogen atoms form hydrogen molecules consisting of 2 protons surrounded by 2 electrons. In the ground state of the hydrogen molecule, both electrons are in the lowest energy standing wave pattern allowed for the molecule. This is the electron cloud we sketched in Figure (19-8) in our discussion of the molecular forces between hydrogen atoms.

The Pauli exclusion principle requires that no two electrons be in exactly the same state. If the two electrons in the hydrogen molecule are in the same standing wave pattern, then they must have opposite spins in order to satisfy the exclusion principle.

If we try to flip one of the electron spins with a radio wave, the spin flipped electron cannot stay in the low energy standing wave pattern, for then we would have two electrons with the same spin in the same wave pattern. In order to flip the spin of one of the electrons, we must supply not only the spin flip energy  $2\mu_B B$ , but also enough energy to raise the electron into a higher energy standing wave pattern. Going to a higher energy standing wave requires much more energy than flipping a spin, thus photons with an energy hf equal to  $2\mu_B B$  will have no effect on hydrogen molecules.

When two electrons are in the same standing wave, we say that the electrons are *paired*. In order to see the spin flip energy, we need a substance with an *unpaired* electron, a substance where the electron spin can be flipped without otherwise disturbing the structure of the substance. Such unpaired electrons can be quite chemically active and are known as *free radicals*. An example of a substance with such an unpaired electron is the crystalline organic chemical diphenyl-picryl-hydrazyl or DPPH for short. Since free radicals cause cancer, when we use this substance in our electron spin resonance experiment, we seal it in a small glass vial to keep from coming in contact with it.

To perform the magnetic resonance experiment, we need to both create the radio waves and detect the energy lost to the electrons being flipped. Both of these steps can be accomplished by placing the glass vial containing the DPPH inside the coil of a resonant LC circuit. The circuit, oscillating at its resonant frequency (f) is the source of the photons of energy hf. Detecting the drain of energy from the coil when hf equals  $2\mu_BB$  is the way we detect the spin flips.

The easiest way to perform the experiment is to get the LC circuit oscillating at some frequency (f), and then change the magnetic field strength B until  $2\mu_B B = hf$ . To detect the loss of energy at this point, we use a specially designed LC circuit that is barely oscillating. The circuit is designed to stop oscillating if any energy is being drained from the circuit. To detect whether or not the circuit is oscillating, another circuit detects the amplitude of the oscillation and puts out a DC voltage proportional to that amplitude. The DC voltage can then be displayed on an oscilloscope.

With this arrangement we can sweep the magnitude of B through the value  $2\mu_B B = hf$  and watch on an oscilloscope the amplitude of the oscillation of the circuit. The result is shown in the solid curve of Figure (2). We see six peaks because the conditions  $2\mu_B B = hf$  was met six times during the 40 milliseconds shown in the diagram.

To produce the magnetic field B, the probe containing the vial of DPPH was placed at the center of our familiar Helmholtz coils as shown in Figure (4). As in our magnetic field mapping experiment *[Figure (24) on page 30-24]*, we power the helmholtz coils with a 60Hz current i(t) to produce a sinusoidally varying magnetic field. This current passes through a  $0.1\Omega$  resistor so that we can measure the strength of the magnetic field by plotting the voltage V(t) across the resistor. The result is the dashed curve seen in Figure (2). We can estimate the strength of B by using i(t) = V(t)/0.1\Omega and then remembering that for these coils the magnetic field B in gauss is about equal to 8i(t).

As the magnetic field went through somewhat more than one cycle in Figure (2), the condition  $2\mu_B B = hf$ 

was met six times producing the six resonant peaks. To see how the condition was met, consider the detailed diagram of the center peaks shown in Figure (3). The first peak, at the time of 18.7 milliseconds, occurred when the magnetic field had a magnitude of 8.4 gauss. [We calculated this from  $105 \times 10^{-3}$  volts  $0.1\Omega$ = 1.05 amps, and then  $B(gauss) = 8i(t) = 8 \times 1.05$  amps = 8.4 gauss.] At time t = 20.1 milliseconds, the magnitude of B goes down through zero, and then reaches a magnitude of -8.4 gauss at a time t = 21.5milliseconds. We get a peak at both +8.4 gauss and -8.4 gauss because the resonance does not depend upon which of the two ways the magnetic field was pointing. Looking back at Figure (2), we see that B had a magnitude of + or -8.4 gauss six times, which is why we got the six peaks.

For the experiment shown in Figures (2) and (3), the photons in the resonant LC circuit, the photons flipping the spin of the DPPH electrons, had a frequency f = 28 megacycles. We can use the fact that these photons flipped the spins when the magnetic field was about 8.4 gauss to calculate our value of the electron magnetic moment  $\mu_{\rm B}$ . We have



The solid curve shows the resonant peaks while the dashed curve is proportional to the strength of the magnetic field.

$$2\mu_{B}B = hf$$
  
 $\mu_{B} = \frac{hf}{2B}$ 

Using the values

$$h = \frac{6.63 \times 10^{-34} \text{ joule sec}}{1.6 \times 10^{-19} \frac{\text{ joules}}{\text{eV}}} = 4.14 \times 10^{-15} \text{ eV sec}$$
$$f = 28 \times 10^6 \left(\frac{1}{\text{sec}}\right)$$
$$B = 8.4 \text{ gauss} \times 10^{-4} \left(\frac{\text{tesla}}{\text{gauss}}\right) = 8.4 \times 10^{-4} \text{tesla}$$

we get

$$\mu_{\rm B} = \frac{4.14 \times 10^{-15} \,\text{eV sec} \times 28 \times 10^{6} \left(\frac{1}{\text{sec}}\right)}{2 \times 8.4 \times 10^{-4} \text{tesla}}$$

 $\mu_{\rm B}. = 6.9 \times 10^{-5} \frac{\rm eV}{\rm tesla} \qquad our \\ result \qquad (17)$ 

This result is nearly 20% above the known value

$$\mu_{\rm B.} = 5.79 \times 10^{-5} \frac{\rm eV}{\rm tesla} \qquad accepted \\ value \qquad (5)$$



#### Figure 4

Electron spin resonance apparatus. The coil containing the vial of DPPH is at the tip of the probe, which is at the center of the Helmholtz coils. The capacitor of the LC circuit, and the controls, are at the other end of the probe. The Helmholtz coils are being driven by a 60Hz alternating current. As a result, the magnitude of the magnetic field sweeps back and forth through the resonant value.



We get a resonant peak for both orientations of the magnetic field.

The source of this error is in our measurement of the strength of the magnetic field. We have relied on the accuracy of the value of the  $0.1\Omega$  resistor through which the helmholtz coil current i(t) passes, and then used the approximate formula B(gauss) = 8i(t). One can obtain much more accurate results using the precision search coil shown in Figure (5). This is a 100 turn coil wound on a 1 inch (2.54cm) plastic rod. Using the techniques discussed in the magnetic mapping experiment, one can accurately relate the voltage  $V_R$  measured across the 0.1 $\Omega$  resistor to the actual value of B. We have encouraged students who wish to do a project involving electron spin resonance to see how accurate a value of  $\mu_B$  they can obtain using this precision search coil.



**Figure 5** 100 turn search coil for accurately determining the magnetic field.

#### **Exercise 6**

In our electron spin resonance experiment, we saw that an electron in a magnetic field had two energy states, and that the difference in the energy between the states was proportional to the strength B of the magnetic field. We measured this energy by placing the electrons in an oscillating electromagnetic field of a given frequency (around 30 megacycles) and observing at what values of the magnetic field we got a transition between the two states.

In Figure (6) we have a somewhat similar situation except the object being studied is a HD (Hydrogen-Deuterium) molecule. In this molecule the Hydrogen nucleus (a proton) weakly interacts with the Deuterium nucleus (a proton and a neutron). These two nuclei form a system with several energy states or levels. If you apply a magnetic field B to the HD molecule, the difference in the energy between the states is related to the strength of B. The energy difference between the states can be measured by applying an oscillating electromagnetic field of a given frequency and observing at what values of the magnetic field we get a transition between states.



H. Benoit and P. Piejus,Compt. Rend. 265B, 101 (1967). "Spectre de R.M.N. de HD á une Fréquence de 54 Hz." Transition from the 3/2,1/2 to the 3/2,3/2 states of the nuclei in an HD molecules at 20 kelvins. The nuclei were pre alligned (polarized) in an 8 killogauss field.

#### Figure 6

NMR data on liquid Hydrogen-Deuterium.

Figure (6) is a nuclear magnetic resonance scan of the HD molecules. As in our electron spin resonance experiment, the molecules are placed in an oscillating electromagnetic field of a given frequency, and the strength of a uniform magnetic field B is varied. Two resonance peaks are observed, but they represent the same transition between the energy levels, since the transition does not depend upon the sign of the magnetic field.

One of the main differences between the electron system we studied in the lab and the HD molecule, is that the energy level splitting is really really small in the HD molecule experiment compared to the splitting we observed in the electron experiment. Instead of fields of tens of gauss and frequencies of around 30 Megacycles, in the HD experiment of Figure (6), the frequency was **54 cycles per second** and the field B was just under **20 milligauss** (.020 gauss or .000002 tesla)! This example demonstrates the enormous range of applicability of the magnetic resonance experiments.

For this exercise, we want you to calculate the splitting between the two energy levels involved in the resonance transitions seen in Figure (6). Give the answer in ergs or joules, and in electron volts. Comment on the reasonableness of your answer—do you think your result is too big, too small, or perhaps OK.

# **APPENDIX**

# CLASSICAL PICTURE OF MAGNETIC INTERACTIONS

At the end of Chapter 31, we discussed the magnetic moment of a current loop, deriving the formulas

 $\vec{A}$  = area of current loop

$$\vec{\mu} \equiv i\vec{A}$$
 magnetic moment (31-34)

$$\vec{t} = \vec{\mu} \times \vec{B}$$
 magnetic torque (31-35)

$$E_{mag} = -\vec{\mu} \cdot \vec{B}$$
 magnetic energy (31-37)

These equations applied to a current loop whose current i and area  $|\vec{A}|$  were unaffected by the magnetic field. We only allowed the magnetic field to change the orientation of the loop via the torque  $\vec{\tau}$ .

We then treated a charged particle in a circular orbit as a current loop. Using the definition  $\vec{\mu} = i\vec{A}$ , we found that the orbiting particle had a magnetic moment  $\vec{\mu}$ related to its angular momentum  $\vec{L}$  by

$$\vec{\mu} = \frac{q}{2m}\vec{L}$$
(31-39)

where q is the charge and m the mass of the particle.

To use the magnetic energy formula  $E_{mag} = -\vec{\mu} \cdot \vec{B}$ , we have to make the same assumptions about the orbiting particle as we did about the current loop. Namely we have to assume that the magnetic field alters only the orientation of the orbit and not the particle's speed v or orbital radius r. Since  $|\vec{L}| = mvr$ , we are thus assuming that the magnetic field does not affect the magnitude of the particle's angular momentum.

For a classical particle, such an argument is not reasonable. If we turn on a magnetic field, we change the magnetic flux through the orbit and thus by Faraday's law induce a voltage around the orbit. This induced voltage should affect both the particle's speed and orbital radius. But if the angular momentum of the particle is quantized, if the magnitude of  $E_{mag} = -\vec{\mu} \cdot \vec{B}$  cannot change, then our current loop analysis and magnetic energy formula  $E_{mag} = -\vec{\mu} \cdot \vec{B}$  has a better chance of working. There is no reason to expect any classical formulas to apply to atomic or subatomic systems. What we are looking for are those that do.

To apply classical formulas to particle spins, we have to fudge the relationship between the particle's magnetic moment  $\vec{\mu}$  and its spin angular momentum. As a general relationship between magnetic moment  $\vec{\mu}$  and angular momentum  $\vec{L}$ , we will rewrite Equation 31-39 in the form

$$\vec{\mu} = g\left(\frac{q}{2m}\vec{L}\right) \tag{A-1}$$

where g is our fudge factor. It is the factor we have to introduce to make classical calculations give the correct results.

The value of g depends upon the kind of system we are talking about. If we are talking about the angular momentum of an electron in orbit about a nucleus, then g=1 and the classical equations work. If we are talking about the spin angular momentum of an electron, then g=2. For electron spin, the classical formulas are off by a factor of 2. For the proton, g has to have the value  $2 \times 2.79$  in order to get the proton magnetic moment given in Equation 9. Since protons and neutrons are composite particles made from quarks it should not be surprising that g should have a peculiar value. This factor g is called either the *gyromagnetic ratio* or *Landé g factor*. Fudge factors sound better if you give them impressive names.

Combining our modified Equation A-1 for  $\vec{\mu}$  with 31-37 for E, we get

$$E_{\text{mag}} = -\vec{\mu} \cdot \vec{B} = g\left(\frac{q}{2m}\right)\vec{L} \cdot \vec{B} \qquad (A-2)$$

as the semi classical formula for the magnetic energy of a particle in a magnetic field. We say semi classical because of the correction factor g. If the particle has an angular momentum  $\vec{L}$  and the magnetic field exerts a torque  $\vec{\tau}$ , the particle should precess like a gyroscope. Thus we can compare a bicycle wheel gyroscope subject to a gravitational torque to a current loop of magnetic moment  $\vec{\mu}$  subject to a magnetic torque. To simplify the analysis, we are assuming that the magnetic moment  $\vec{\mu}$  lies in the plane perpendicular to  $\vec{B}$  so the magnetic torque  $\vec{\tau} = \vec{\mu} \times \vec{B}$  has a magnitude  $\mu B$ .

Following the standard analysis of a gyroscope, we predict that the magnetic moment vector  $\vec{\mu}$  should precess about the magnetic field vector  $\vec{B}$  at a rate  $\omega_{\text{precession}}$  given by

$$\omega_{\text{precession}} = \frac{\tau}{L} = \frac{\mu B}{L}$$
(12-58)

a result we derived back in Chapter 12. Using our semi classical formula A1 for  $\mu$ , we get

$$\omega_{\text{precession}} = g\left(\frac{qL}{2m}\right)\frac{B}{L}$$

The L's cancel, and we are left with

$$\omega_{\text{precession}} = g\left(\frac{q}{2m}\right)B$$
 (A-3)

The quantity  $\omega_{\text{precession}}$  is the precessional frequency in radians per second. To convert this to cycles per second, we divide by  $2\pi$  radians/cycle to get

$$f_{\text{precession}} = \frac{\omega_{\text{precession}}}{2\pi} = \frac{g}{2\pi} \left(\frac{q}{2m}\right) B$$
 (A-4)

as the precessional frequency of a charged orbiting or spinning particle in a magnetic field.

Applying Equation A-4 to the spin of a particle, we predict classically that if the particle is subject to an alternating electric and magnetic fields of frequency f, there will be a resonance and the particle can gain magnetic energy if the frequency f equals the precessional frequency  $f_{precession}$ .

To go to the quantum picture, multiply Equation A-4 through by Planck's constant h, to get

$$hf_{\text{precession}} = g \frac{h}{2\pi} \frac{q}{2m} B$$
$$= g \left(\frac{q\hbar}{2m}\right) B \qquad (A-5)$$

Applying this to an electron spin, setting q and m to the charge and mass of an electron, we get

$$hf_{precession} = g\mu_B B \qquad \begin{array}{c} classical theory \\ with factor g \end{array}$$
 (A-6)

where  $\mu_{\rm B} = e\hbar/2m$  is the Bohr magneton.

In the quantum picture, the electron gains magnetic energy if the photons in the radio wave have the right amount of energy to flip the spin of the electron. The Dirac equation gave the spin flip energy as

$$\Delta E = 2\mu_{\rm B}B \qquad \begin{array}{l} energy \ required \\ to \ flip \ the \\ electron \ spin \end{array} (3 \ repeated)$$

In Equation 7, we equated this energy to the photon energy to get

hf = 
$$2\mu_B B$$
 set spin flip  
energy equal to (7 repeated)  
photon energy

as the formula giving the frequency of the radio wave that can add magnetic energy to the electron. Comparing Equation (7) with the semi classical result (A-6), we see that the Landé g factor, the gyromagnetic ratio g, has to be set equal to 2 for the semi classical theory to agree with the Dirac equation

$$g = 2 \qquad \begin{array}{c} gyromagnetic ratio\\ for the electron spin \end{array}$$
(A-7)

# Chapter 40 Quantum Mechanics

That light had both a particle and a wave nature became apparent with Einstein's explanation of the photoelectric effect in 1905. One might expect that such a discovery would lead to a flood of publications speculating on how light could behave both as a particle and a wave. But no such response occurred. The particle wave nature was not looked at seriously for another 18 years, when de Broglie proposed that the particle wave nature of the electron was responsible for the quantized energy levels in hydrogen. Even then there was great reluctance to accept de Broglie's proposal as a satisfactory thesis topic.

Why the reluctance? Why did it take so long to deal with the particle-wave nature, first of photons then of electrons? What conceptual problems do we encounter when something behaves both as a particle and as a wave? How are these problems handled? That is the subject of this chapter.

# TWO SLIT EXPERIMENT

Of all the experiments in physics, it is perhaps the 2 slit experiment that most clearly, most starkly, brings out the problems encountered with the particle-wave nature of matter. For this reason we will use the 2 slit experiment as the basis for much of the discussion of this chapter.

Let us begin with a review of the 2 slit experiment for water and light waves. Figure (1) shows the wave pattern that results when water waves emerge from 2 slits. The lines of nodes are the lines along which the waves from one slit just cancel the waves coming from the other. Figure (2) shows our analysis of the 2 slit pattern. The path length difference to the first minimum must be half a wavelength  $\lambda/2$ . This gives us the two similar triangles shown in Figure (2). If  $y_{min}$  is



**Figure 1** *Water waves emerging from two slits.* 



Figure 2

Analysis of the two slit pattern. We get a minimum when the path length difference is half a wavelength.

much less than D, which it is for most 2 slit experiments, then the hypotenuse of the big triangle is approximately D and equating corresponding sides of the similar triangles gives us the familiar relationship

$$\frac{\lambda/2}{d} = \frac{y_{\min}}{D}$$

$$\lambda = \frac{2y_{\min}d}{D}$$
(1)

Figure (3a) is the pattern we get on a screen if we shine a laser beam through 2 slits. To prove that the dark bands are where the light from one slit cancels the light from the other, we have in Figure (3b) moved a razor blade in front of one of the slits. We see that the dark bands disappear, and we are left with a one slit pattern. The dark bands disappear because there is no longer any cancellation of the waves from the 2 slits.



Two slit interference pattern for light. The closely spaced dark bands are where the light from one slit cancels the light from the other.



#### Figure 3b

Move a razor blade in front of one of the slits, and the closely spaced dark bands disappear. There is no more cancellation.

In 1961, Claus Jönsson did the 2 slit experiment using electrons instead of light, with the results shown in Figure (4). Assuming that the electron wavelength is given by the de Broglie formula  $p = h/\lambda$ , the dark bands are located where one would expect waves from the 2 slits to cancel. The 2 slit experiment gives the same result for light and electron waves.

# The Two Slit Experiment from a Particle Point of View

In Figure (3a), the laser interference patterns were recorded on a photographic film. The pattern is recorded when individual photons of the laser light strike individual silver halide crystals in the film, producing a dark spot where the photon landed. Where the image shows up white in the positive print, many photons have landed close together exposing many crystal grains.

In a more modern version of the experiment one could use an array of photo detectors to count the number of photons landing in each small element of the array. The number of counts per second in each detector could then be sent to a computer and the image reconstructed on the computer screen. The result would look essentially the same as the photograph in Figure (3a).

The point is that the image of the two slit wave pattern for light is obtained by counting particles, not by measuring some kind of a wave height. When we look at the two slit experiment from the point of view of counting particles, the experiment takes on a new perspective.



**Figure 4** *Two slit experiment using electrons. (By C. Jönsson)* 

Imagine yourself shrunk down in size so that you could stand in front of a small section of the photographic screen in Figure (3a). Small enough that you want to avoid being hit by one of the photons on the laser beam. As you stand at the screen and look back at the slits, you see photons being sprayed out of both slits as if two machine guns were firing bullets at you, but you discover that there is a safe place to stand. There are these dark bands where the particles fired from one slit cancel the particles coming from the other.

Then one of the slits is closed, there is no more cancellation, the dark bands disappear as seen in Figure (3b). There is no safe place to stand when particles are being fired at you from only one slit. It is hard to imagine in our large scale world how it would be safe to have two machine guns firing bullets at you, but be lethal if only one is firing. It is hard to visualize how machine gun bullets could cancel each other. But the particle wave nature of light seems to require us to do so. No wonder the particle nature of light remained an enigma for nearly 20 years.

# Two Slit Experiment—One Particle at a Time

You might object to our discussion of the problems involved in interpreting the two slit experiment. After all, Figure (1) shows water waves going through two slits and producing an interference pattern. The waves from one slit cancel the waves from the other at the lines of nodes. Yet water consists of particles—water molecules. If we can get a two slit pattern for water molecules, what is the big deal about getting a two slit pattern for photons ? Couldn't the photons somehow interact with each other the way water molecules do, and produce an interference pattern?

Photons do not interact with each other the way water molecules do. Two laser beams can cross each other with no detectable interaction, while two streams of water will splash off of each other. But one still might suspect that the cancellation in the two slit experiment for light is caused by some kind of interaction between the photons. This is even more likely in the case of electrons, which are strongly interacting charged particles.



a) 10 dots

b) 100 dots



c) 1000 dots



d) 10000 dots



**Predicted pattern** 



Experimental results by C. Jönsson

#### Figure 5

Computer simulation of the 2 slit electron diffraction experiment, as if the electrons had landed one at a time. In an earlier text, we discussed the possibility of an experiment in which electrons would be sent through a two slit array, one electron at a time. The idea was to eliminate any possibility that the electrons could produce the two slit pattern by bouncing into each other or interacting in any way. Since the experiment had not yet been done, we drew a sketch of what the results should look like. That sketch now appears in a number of introductory physics texts.

When he saw the sketch, Lawrence Campbell of the Los Alamos Scientific Laboratories did a computer simulation of the experiment. We will first discuss Campbell's simulation, and then compare the simulation with the results of the actual experiment which was performed in 1991.

It is not too hard to guess some of the results of sending electrons through two slits, one at a time. After the first electron goes through you end up with one dot on the screen showing where the electron hit. The single dot is not a wave pattern. After two electrons, two dots; you cannot make much of a wave pattern out of two dots.

If, after many thousands of electrons have hit the screen, you end up with a two slit pattern like that shown in Figure (4), that means that none of the electrons land where there will eventually be a dark band. You know where the first dot, and the second dot cannot be located. Although two dots do not suggest a wave pattern, some aspects of the wave have already imposed themselves by preventing the dots from being located in a dark band.

To get a better idea of what is happening, let us look at Campbell's simulation in Figure (5). In (5a), and (5b) we see 10 dots and 100 dots respectively. In neither is there an apparent wave pattern, both look like a fairly random scatter of dots. But by the time there are 1000 dots seen in (5c), a fairly distinctive interference pattern is emerging. With 10,000 dots of (5d), we see a fairly close resemblance between Campbell's simulation and Jönsson's experimental results. Figure (5e) shows the wave pattern used for the computer simulation.

Although the early images in Figure (5) show nearly random patterns, there must be some order. Not only do the electrons not land where there will be a dark band, but they must also accumulate in greater numbers where the brightest bands will eventually be. If this were a roulette type of game in Las Vegas, you should put your money on the center of the brightest band as being the location most likely to be hit by the next electron.

Campbell's simulation was done as follows. Each point on the screen was assigned a probability. The probability was set to zero at the dark bands and to the greatest value in the brightest band. Where each electron landed was randomly chosen, but a randomness governed by the assigned probability.

How to assign a probability to a random event is illustrated by a roulette wheel. On the wheel, there are 100 slots, of which 49 are red, 49 black and 2 green. Thus where the ball lands, although random, has a 49% chance of being on red, 49% on black, 2% on green, and 0% on blue, there being no blue slots.

In the two slit simulation, the probability of the electron landing at some point was proportional to the intensity of the two slit wave pattern at that point. Where the wave was most intense, the electron is most likely to land. Initially the pattern looks random because the electrons can land with roughly equal probability in any of the bright bands. But after many thousands of electrons have landed, you see the details of the two slit wave pattern. The dim bands are dimmer than the bright ones because there was a lower probability that the electron could land there.

Figure (6) shows the two slit experiment performed in 1991 by Akira Tonomura and colleagues. The experiment involved a novel use of a superconductor for the two slits, and the incident beam contained so few electrons per second that no more than one electron was between the slits and the screen at any one time. The screen consisted of an array of electron detectors which recorded the time of arrival of each electron in each detector. From this data the researchers could reconstruct the electron patterns after 10 electrons (6a), 100 electrons (6b), 3000 electrons (6c), 20,000 electrons (6d) and finally after 70,000 electrons in Figure (6e). Just as in Campbell's simulation, the initially random looking patterns emerge into the full two slit pattern when enough electrons have hit the detectors.



#### Figure 6

Experiment in which the 2 slit electron interference pattern is built up one electron at a time. (A. Tonomura, J. Endo, T. Matsuda, T. Kawasaki, American Journal of Physics, Feb. 1989. See also Physics Today, April 1990, Page 22.)

# **Born's Interpretation** of the Particle Wave

In 1926, while calculating the scattering of electron waves, Max Born discovered an interpretation of the electron wave that we still use today. In Born's picture, the electron is actually a particle, but it is the electron wave that governs the behavior of the particle. The electron wave is a probability wave governing the probability of where you will find the electron.

To apply Born's interpretation to the two slit electron experiment, we do what Campbell did in the simulation of Figure (5). We first calculate what the wave pattern at the screen would be for a wave passing through the two slits. It is the two slit interference pattern we have seen for water waves, light waves and electron waves. We then interpret the intensity of the pattern at some point on the screen as being proportional to the probability that the electron will land at that point. We cannot predict where any given electron will actually land, any more than we can predict where the ball will end up on the roulette wheel. But we can predict what the pattern will look like after many electrons have landed. If we repeat the experiment, the electrons will not land in the same places, but eventually the same two slit pattern will result.

#### **Exercise 1**

Figure (36-16) reproduced here, shows the diffraction pattern produced when a beam of electrons is scattered by the atoms of a graphite crystal. Explain what you would expect to see if the electrons went through the graphite crystal one at a time and you could watch the pattern build up on the screen. Could you market this apparatus in Las Vegas, and if so, how would you use it?

Figure 36-16

Diffraction pattern produced by electrons passing through a graphite crystal.



# **Photon Waves**

Both electrons and photons have a particle-wave nature related by the de Broglie formula  $p = h/\lambda$ , and both produce a two slit interference pattern. Thus one would expect that the same probability interpretation should apply to electron waves and light waves.

We have seen, however, that a light wave, according to Maxwell's equations, consists of a wave of electric and magnetic fields  $\vec{E}$  and  $\vec{B}$ . These are vector fields that at each point in space have both a magnitude and a direction. Since probabilities do not point anywhere, we cannot directly equate  $\vec{E}$  and  $\vec{B}$  to some kind of probability.

To see how to interpret the wave nature of a photon, let us first consider something like a radio wave or a laser beam that contains many billions of photons. In our discussion of capacitors in Chapter 27, we saw that the energy density in a classical electric field was given by

$$\frac{\text{Energy}}{\text{density}} = \frac{\varepsilon_0 E^2}{2} \quad \frac{\text{energy density in}}{\text{an electric field}} \quad (27-36)$$

where  $E^2 = \vec{E} \cdot \vec{E}$ . In an electromagnetic wave there are equal amounts of energy in the electric and the magnetic fields. Thus the energy density in a classical electromagnetic field is twice as large as that given y Equation 27-36, and we have

$$E \frac{\text{joules}}{\text{meter}^3} = \varepsilon_0 E^2$$
 energy density in an   
electricomagnetic wave

If we now picture the electromagnetic wave as consisting of photons whose energy is given by Einstein's photoelectric formula

$$E_{photon} = hf \frac{joules}{photon}$$

then the density of photons in the wave is given by

$$n = \frac{\varepsilon_0 E^2 \text{ joules / meter}^3}{\text{hf joules / photon}}$$

$$n = \frac{\varepsilon_0 E^2}{hf} \frac{\text{photons}}{\text{meter}^3} \qquad \begin{array}{c} \text{density of photons} \\ \text{in an electromagnetic} \\ \text{wave of frequency } f \end{array} (1)$$

where f is the frequency of the wave.

In Exercise 2, we have you estimate the density of photons one kilometer from the antenna of the student AM radio station at Dartmouth College. The answer turns out to be around .25 billion photons/cc, so many photons that it would be hard to detect them individually.

#### **Exercise 2**

To estimate the density of photons in a radio wave, we can, instead of calculating  $\vec{E}$  for the wave, simply use the fact that we know the power radiated by the station. As an example, suppose that we are one kilometer away from a 1000 watt radio station whose frequency is  $1.4 \times 10^{6}$ Hz. A 1000 watt station radiates 1000 joules of energy per second or  $10^{-6}$  joules in a nanosecond. In one nanosecond the radiated wave moves out one foot or about 1/3 of a meter. If we ignore spatial distortions of the wave, like reflections from the ground, etc., then we can picture this  $10^{-6}$  joules of energy as being located in a spherical shell 1/3 of a meter thick, expanding out from the antenna.

(a) What is the total volume of a spherical shell 1/3 of a meter thick and 1 kilometer in radius?

(b) What is the average density of energy, in joules/m<sup>3</sup> of the radio wave 1 kilometer from the antenna

(c) What is the energy, in joules, of one photon of frequency  $1.4 \times 10^{6}$  Hz?

(d) What is the average density of photons in the radio wave 1 kilometer from the station? Give the answer first in photons/m<sup>3</sup> and then photons per cubic centimeter. (The answer should be about .25 billion photons/cm<sup>3</sup>.)

Now imagine that instead of being one kilometer from the radio station, you were a million kilometers away. Since the volume of a spherical shell 1/3 of a meter thick increases as  $r^2$ , (the volume being  $(1/3) \times 4\pi r^2$ ) the density of photons would decrease as  $1/r^2$ . Thus if you were  $10^6$  times as far away, the density of photons would be  $10^{-12}$  times smaller. At one million kilometers, the average density of photons in the radio wave would be

number of photons  
per cubic centimeter  
at 1 million kilometers
$$\begin{cases}
= \frac{\text{number at 1km}}{10^{12}} \\
= \frac{.25 \times 10^9}{10^{12}} \\
= .00025 \frac{\text{photons}}{\text{cm}^3} \quad (2)
\end{cases}$$

In the classical picture of Maxwell's equations, the radio wave has a continuous electric and magnetic field even out at 1 million kilometers. You could calculate the value of  $\vec{E}$  and  $\vec{B}$  out at this distance, and the result would be sinusoidally oscillating fields whose structure is that shown back in Figure (32-23). But if you went out there and tried to observe something, all you would find is a few photons, on the order of .25 per liter (about one per gallon of space). If you look in 1 cubic centimeter of space, chances are you would not find a photon.

So how do you use Maxwell's equations to predict the results of an experiment to detect photons a million kilometers from the antenna? First you use Maxwell's equation to calculate  $\vec{E}$  at the point of interest, then evaluate the quantity ( $\epsilon_0 \vec{E} \cdot \vec{E} / hf$ ), and finally interpret the result as the probability of finding a photon in the region of interest. If, for example, we were looking in a volume of one cubic centimeter, the probability of finding a photon there would be about .00025 or .025%.

This is an explicit prescription for turning Maxwell's theory of electromagnetic radiation into a probability wave for photons. If the wave is intense, as it was close to the antenna, then  $(\epsilon_0 \vec{E} \cdot \vec{E}/hf)$  represents the density of photons. If the wave is very faint, then  $(\epsilon_0 \vec{E} \cdot \vec{E}/hf)$  becomes the probability of finding a photon in a certain volume of space.

# **Reflection and Fluorescence**

An interesting example of the probability interpretation of light waves is provided by the phenomena of reflection and of fluorescence.

When a light beam is reflected from a metal surface, the angle of reflection, labeled  $\theta_r$  in Figure (7a) is equal to the angle of incidence  $\theta_i$ . The reason for this is seen in Figure (7b). The incident light wave is scattered by many atoms in the metal surface. The scattered waves add up to produce the reflected wave as shown in Figure (7b). Any individual photon in the incident wave must have an equal probability of being scattered by all of these atoms in order that the scattered probability waves add up to the reflected wave shown in (7b).

When you have a fluorescent material, you see a rather uniform eerie glow rather than a reflected wave. The light comes out in all directions as in Figure (8a).



**Figure 7a** When a light wave strikes a mirror, the angle of incidence equals the angle of reflection.



#### Figure 7b

The reflected wave results from the scattering of the incident wave by many atoms. If the incident wave contains a single photon, that photon must have an equal probability of being scattered by many atoms in order to emerge in the reflected wave. The wavelength of the light from a fluorescent material is not the same wavelength as the incident light. What happens is that a photon in the incident beam strikes and excites an individual atom in the material. The excited atom then drops back down to the ground state radiating two or more photons to get rid of the excitation energy. (Ultraviolet light is often used in the incident beam, and we see the lower energy visible photons radiated from the fluorescing material.)

The reason that fluorescent light emerges in many directions rather than in a reflected beam is that an individual photon in the incident beam is absorbed by and excites one atom in the fluorescent material. There is no probability that it has struck any of the other atoms. The fluorescent light is then radiated as a circular wave from that atom, and the emerging photon has a more or less equal probability of coming out in all directions above the material.



Figure 8a

When a beam of light strikes a fluorescent material, we see an eerie glow rather than a normal reflected light.



#### Figure 8b

Fluorescence occurs when an individual atom is excited and radiates its extra energy as two distinct photons. Since there is no chance that the radiation came from other atoms, the radiated wave emerges only from the excited atom.

# A Closer Look at the Two Slit Experiment

While the probability interpretation of electron and photon waves provides a reasonable explanation of some phenomena, the interpretation is not without problems. To illustrate what these problems are, consider the following thought experiment.

Imagine that we have a large box with two slits at one end and a photographic film at the other, as shown in Figure (9). Far from the slits is an electron gun that produces a weak beam of electrons, so weak that on the average only one electron per hour passes through the slits and strikes the film. For simplicity we will assume that the electrons go through the slits on the hour, there being the 9:00 AM electron, the 10:00 AM electron, etc.

The electron gun is one of the simple electron guns we discussed back in Chapter 28. The beam is so spread out that there is no way it can be aimed at one slit or the other. Our beam covers both slits, meaning that each of the electrons has an equal chance of going through the top or bottom slit.

We will take the probability interpretation of the electron wave seriously. If the electron has an equal probability of passing through either slit, then an equally intense probability wave must emerge from both slits. When the probability waves get to the photographic film, there will be bands along which waves from one slit cancel waves from the other, and we should eventually build up a two slit interference pattern on the film.

Suppose that on our first run of the thought experiment, we do build up a two slit pattern after many hours and many electrons have hit the film. We will now repeat the experiment with a new twist. We ask for a volunteer to go inside the box, look at the slits, and see which one each electron went through. John volunteers, and we give him a sheet of paper to write down the results. To make the job easier, we tell him to just look at the bottom slit on the hour to see if the electron went through that slit. If for example he sees an electron come out of the bottom slit at 9:00 AM, then the 9:00 AM electron went through the bottom slit. If he saw no electron at 10:00 AM, then the 10:00 AM electron must have gone through the upper slit.

If John does his job carefully, what kind of a pattern should build up on the film after many electrons have gone through? If the 9:00 AM electron was seen to pass through the bottom slit, then there is no probability that it went through the top slit. As a result, a probability wave can emerge only from the bottom slit, and there can be no cancellation of probability waves at the photographic film. Since the 10:00 AM electron did not go through the bottom slit, the probability wave must have emerged only from the top slit and there again can be no cancellation of waves at the photographic film.

If John correctly determines which slit each electron went through, there can be no cancellation of waves from the two slits, and we have to end up with a one slit pattern on the film. Just the knowledge of which slit each electron went through has to change the two slit pattern into a one slit pattern. With Born's probability interpretation of electron waves, just the *knowledge* of which slit the electrons go through *changes the result of the experiment*. Does this really happen, or have we entered the realm of metaphysics?



#### Figure 9

In this thought experiment, we consider the possibility that someone is looking at the two slits to see which slit each electron comes through.

Let us return to our thought experiment. John has been in the box for a long time now, so that a number of electrons have hit the film. We take the film out, develop it, and clearly see a two slit interference pattern emerging. There are the dark bands along which waves from one slit cancel the waves from the other slit.

Then we go over to the door on the side of the box, open it and let John out, asking to see his results. We look at his sheet of paper and nothing is written on it. "What were you doing all of that time," we ask. "What do you mean, what was I doing? How could I do anything? You were so careful sealing up the box from outside disturbances that it was dark inside. I couldn't see a thing and just had to wait until you opened the door. Not much of a fun experiment."

"Next time," John said, "give me a flashlight so I can see the electrons coming through the slits. Then I can fill out your sheet of paper."

"Better be careful," Jill interrupts, "about what kind of a flashlight you give John. A flashlight produces a beam of photons, and John can only see a passing electron if one of the flashlight's photons bounces off the electron."

"Remember that the energy of a photon is proportional to its frequency. If the photons from John's flashlight have too high a frequency, the photon hitting the passing electron will change the motion of the electron and mess up the two slit pattern. Give John a flashlight that produces low frequency, low energy photons, so he won't mess up the experiment." "But," Bill responds, "a low frequency photon is a long wavelength photon. Remember that demonstration where waves were scattered from a tiny object? The scattered waves were circular, and contained no information about the shape of the object (Figure 36-1). You can't use waves to study details that are much smaller than the wavelength of the wave. That is why optical microscopes can't be used to study viruses that are smaller than a wavelength of visible light."

"If John's flashlight," Bill continues, "produced photons whose wavelength was longer than the distance between the two slits, then even if he hit the electron with one of the photons in the wave, John could not tell which slit the electron came through."

"Let us do some calculations," the professor says. "The most delicate way we can mess up the experiment is to hit an electron sideways, changing the electron's direction of motion so that if it were heading toward a maxima, it will instead land in a minima, filling up the dark bands and making the pattern look like a one slit pattern. Here is a diagram for the situation (Figure 10)."

"In the top sketch (10a), John is shining his flashlight at an electron that has just gone through the slit and is heading toward the central maximum. In the middle sketch (10b) the photon has knocked the electron sideways, so that it is now headed toward the first minimum in the diffraction pattern. Let us assume that all the photon's momentum  $\vec{p}_{photon}$  has been transferred to the electron, so that the electron's new momentum is now

**Figure 36-1 (reproduced)** *If an object is smaller than a wavelength, the scattered waves are circular and do not contain information about the shape of the object.* 



Incident and scattered wave



After incident wave has passed



Figure 10a

In order to see the electron, John uses a flashlight, and strikes the electron with a photon.



#### Figure 10b

Assume the photon's momentum has been absorbed by the electron. This could deflect the electron's path by an angle  $\theta$ .



#### Figure 10c

Analysis of the two slit pattern. The angle to the first minimum is determined by using similar triangles. If the angle  $\theta$  is small, then  $\sin \theta \approx \theta$ .

$$\vec{p}_{electron-new} = \vec{p}_{electron-old} + \vec{p}_{photon}$$
 (3)

The angle  $\theta$  by which the electron is deflected is approximately given by

$$\theta \approx \frac{p_{\text{photon}}}{p_{\text{electron}}} = \frac{h/\lambda_{\text{photon}}}{h/\lambda_{\text{electron}}}$$
$$\theta \approx \frac{\lambda_{\text{electron}}}{\lambda_{\text{photon}}}$$
(4)

where we used the de Broglie formula for the photon and electron momenta."

"In the bottom sketch we have the usual analysis of a two slit pattern. If the angle  $\theta$  to the first minimum is small, which it usually is for a two slit experiment, then by similar triangles we have

$$\theta \approx \frac{y_{\min}}{D} = \frac{\lambda_{electron}/2}{d}$$
 (5)

Equating the values of  $\theta$  from equations (5) and (4), we get

$$\theta = \frac{\lambda_{\text{electron}}}{2d} = \frac{\lambda_{\text{electron}}}{\lambda_{\text{photon}}}$$
(6)

"Look!" Bill says, " $\lambda_{electron}$  cancels and we are left with "

$$\lambda_{\text{photon}} = 2d \tag{7}$$

"I told you," Jill interrupts, "that you had to be careful about what wavelength photons John could use. Here we see that if John's photons have a wavelength of 2d or less, his photons will carry enough of a punch, enough momentum to destroy the two slit pattern. Be sure John's photons have a wavelength longer than 2d so that they will be incapable of knocking an electron from a maxima to a minima."

"No way," responds Bill. "A wavelength of 2d is already too big. John cannot use photons with a wavelength any greater than the slit width d if he wants to see which slit the electron went through. And you want him to use photons with a wavelength greater than 2d!" "That's the dilemma," the professor replies. "If John uses photons whose wavelength is short enough to see which slit the electron went through, he is likely to mess up the experiment and destroy the two slit pattern."

"It looks like the very act of getting information is messing up the experiment," Jill muses.

"It messes it up if we use photons," Bill responds." Let us work out a better experiment where we do a more delicate measurement to see which slit the electron went through. Do the experiment so delicately that we do not affect the motion of the electron, but accurately enough to see which slit the electron went through."

"How would you do that?" Jill asks.

"Maybe I would put a capacitor plate on one of the slits," Bill responds, "and record the capacitor voltage. If the electron went through that slit, the electric field of the electron should affect the voltage on the capacitor and leave a blip on my oscilloscope screen. If I don't see a blip, the electron went through the other slit."

"Would this measurement affect the motion of the electron?" Jill asks.

"I don't see why," Bill responds.

"Think about this," the professor interrupts. "We are now interpreting the electric and magnetic fields of a light wave as a probability wave for photons. In this view, all electric and magnetic phenomena are ultimately caused by photons. The electric and magnetic fields we worked with earlier in the course are now to be thought of as a way of describing the behavior of the underlying photons."

"That's crazy," Bill argues. "You mean, for example, that the good old  $1/r^2$  Coulomb force law that holds the hydrogen atom together, is caused by photons? I don't see how."

"It's hard to visualize, but you can use a photon picture to explain every detail of the interaction between the electron and proton in the hydrogen atom. That calculation was actually done back in 1947. The modern view is that all electric and magnetic phenomena are caused by photons."

"If all electric and magnetic phenomena are caused by photons," Jill observes, "then Bill's capacitor plate and voltmeter, which uses electromagnetic phenomena, is based on photons. Since photons obey the de Broglie relationship, the photons in Bill's experiment should have the same effect as the photons from John's flashlight. If John's photons mess up the experiment, Bill's should too!"

"I have an idea," Bill says. "Aren't there such a thing as gravitational waves?"

"Yes," replies the professor. "They are very hard to make, and very hard to detect. We have not been able to make or detect them yet in the laboratory. But back in the 1970's Joe Taylor at the University of Massachusetts discovered a pair of binary neutron stars orbiting about each other. Since the stars eclipse each other, Taylor could accurately measure the orbital period."

"According to Einstein's theory of gravity, the orbiting neutron stars should radiate gravitational waves and lose energy. Joe Taylor has conclusively shown that the pair of stars are losing energy just as predicted by Einstein's theory. Taylor got the Nobel prize for this work in 1993."

"Is Einstein's theory a quantum theory?" Bill asks.

"What do you mean by that?" Jill asks.

"I mean," Bill responds, "in Einstein's theory, do gravitational waves have a particle wave nature like electromagnetic waves? Are there particles in a gravitational wave like there are photons in a light wave?"

"Not in Einstein's theory," the professor replies. "Einstein's theory is strictly a classical theory. No particles in the wave." "Then if Einstein's theory is correct," Bill continues, "I should be able to make a gravitational wave with a very short wavelength and very little energy."

"Couldn't I then use this short wavelength, low energy, gravitational wave to see which slit the electron went through? I would make the wavelength much shorter than the slit spacing d so that there would be no doubt about which slit the electron went through. But I would use a very low energy, delicate wave so that I would not affect the motion of the electron."

"You could do that if Einstein's theory is right," the professor replies.

"But," Bill responds, "that allows me to tell which slit the electron went through without destroying the two slit pattern. What happens to the probability interpretation of the electron wave? If I know which slit the electron went through, the probability wave must have come from that slit, and we must get a one slit pattern. If John used gravitational waves instead of light waves in his flashlight, he could observe which slit the electron went through without destroying the two slit pattern."

"You have just stumbled upon one of the major outstanding problems in physics," the professor replies. "As far as we know there are four basic forces in nature. They are gravity, the electromagnetic force, the weak interaction, and the so-called gluon force that holds quarks together. Ilisted these in the order in which they were discovered."

"Now three of these forces, all but gravity, are known to have a particle-wave nature like light. All the particles obey the de Broglie relation  $p = h/\lambda$ ."

"As a result, if we perform our two slit electron experiment, trying to see which slit the electron went through, and we use apparatus based on non gravitational forces, we run into the same problem we had with John's flashlight. The only chance we have for detecting which slit the electron went through without messing up the two slit pattern, is to use gravity." "Could Einstein be wrong?" Jill asks. "Couldn't gravitational waves also have a particle nature? Couldn't the gravitational particles also obey the de Broglie relation?"

"Perhaps," the professor replies. "For years, physicists have speculated that gravity should have a particlewave nature. They have even named the particle -- they call it a *graviton*. One problem is that gravitons should be very, very hard to detect. The only way we know that gravitational waves actually exist is from Joe Taylor's binary neutron stars. There are various experiments designed to directly observe gravitational waves, but no waves have yet been seen in these experiments."

"In the case of electromagnetism, we saw electromagnetic radiation -- i.e., light -- long before photons were detected in Hertz's photoelectric effect experiment. After gravitational waves are detected, then we will have to do the equivalent of a photoelectric effect experiment for gravity in order to see the individual gravitons. The main problem here is that the gravitational radiation we expect to see, like that from massive objects such as neutron stars, is very low frequency radiation. Thus we would be dealing with very low energy gravitons which would be hard to detect individually."

"And there is another problem," the professor continues, "no one has yet succeeded in constructing a consistent quantum theory of gravity. There are mathematical problems that have yet to be overcome. At the present time, the only consistent theory of gravity we have is Einstein's classical theory."

"It looks like two possibilities," Jill says. "If the probability interpretation of electron waves is right, then there has to be a quantum theory of gravity, gravitons have to exist. If Einstein's classical theory is right, then there is some flaw in the probability interpretation."

"That is the way it stands now," the professor replies.
# THE UNCERTAINTY PRINCIPLE

We have just seen that, for the probability interpretation of particle-waves to be a viable theory, there can be no way we can detect which slit the electron went through without destroying the two slit pattern. Also we have seen that if every particle and every force have a particle wave nature obeying the de Broglie relationship  $\lambda = h/p$ , then there is no way we can tell which slit the electron went through without destroying the two slit pattern. Both the particle-wave nature of matter, and the probability interpretation of particle waves, lead to a basic limitation on our ability to make experimental measurements. This basic limitation was discovered by Werner Heisenberg shortly before Schrödinger developed his wave equation for electrons. Heisenberg called this limitation the uncertainty principle.

When you cannot do something, when there is really no way to do something, physicists give the failure a name and call it a basic law of physics. We began the text with the observation that you cannot detect uniform motion. Michaelson and Morley thought they could, repeatedly tried to do so, and failed. This failure is known as the principle of relativity which Einstein used as the foundation of his theories of relativity. Throughout the text we have seen the impact of this simple idea. When combined with Maxwell's theory of light, it implied that light traveled at the same speed relative to all observers. That implied moving clocks ran slow, moving lengths contracted, and the mass of a moving object increased with velocity. This led to the relationship  $E = mc^2$  between mass and energy, and to the connections between electric and magnetic fields. The simple idea that you cannot measure uniform motion has an enormous impact on our understanding of the way matter behaves.

Now, with the particle-wave nature of matter, we are encountering an equally universal restriction on what we can measure, and that restriction has an equally important impact on our understanding of the behavior of matter. Our discussion of the uncertainty principle comes at the end of the text rather than at the beginning only because it has taken a while to develop the concepts we need to explain this restriction. With the principle of relativity we could rely on the student's experience with uniform motion, clocks and meter sticks. For the uncertainty principle, we need some understanding of the behavior of particles and waves, and as we shall see, Fourier analysis plays an important role.

There are two forms of the uncertainty principle, one related to measurements of position and momentum, and the other related to measurements of time and energy. They are not separate laws, one can be derived from the other. The choice of which to use is a matter of convenience. Our discussion of the two slit experiment and the de Broglie relationship naturally leads to the position-momentum form of the law, while Fourier analysis naturally introduces the time-energy form.

# POSITION-MOMENTUM FORM OF THE UNCERTAINTY PRINCIPLE

In our two slit thought experiment, in the attempt to see which slit the electron went through, we used a beam of photons whose momenta was related to their wavelength by  $p = h/\lambda$ . The wave nature of the photon is important because we cannot see details smaller than a wavelength  $\lambda$  when we scatter waves from an object. When we use waves of wavelength  $\lambda$ , the uncertainty in our measurement is at least as large as  $\lambda$ . Let us call the uncertainty in the position measurement  $\Delta x$ .

However when we use photons to locate the electron, we are slugging the electron with particles, photons of momentum  $p_{photon} = h/\lambda$ . Since we do not know where the photons are within a distance  $\lambda$ , we do not know exactly how the electron was hit and how much momentum it absorbed from the photon. The electron could have absorbed the full photon momentum  $p_{photon}$  or none of it. If we observe the electron, we make the electron's momentum uncertain by an amount at least as large as  $p_{photon}$ . Calling the uncertainty in the electron's momentum  $\Delta p_{electron}$  we have

$$\Delta p_{\text{electron}} = p_{\text{photon}} = \frac{h}{\lambda} = \frac{h}{\Delta x}$$
 (8)

multiplying through by  $\Delta x$  gives

$$\Delta p \Delta x = h \tag{9}$$

In Equation 9,  $\Delta p$  and  $\Delta x$  represent the smallest possible uncertainties we can have when measuring the position of the electron using photons. To allow for the fact that we could get much greater uncertainties using poor equipment or sloppy techniques, we will write the equation in the form

$$\Delta p\Delta x \ge h \qquad position-momentum form of the uncertainty principle \qquad (10)$$

indicating that the product of the uncertainties is at least as large as Planck's constant h. If all forces have a particle nature, and all particles obey the de Broglie relationship, then the fact that we derived Equation 10 using photons makes no difference. We have to get the same result using any particle, in any possible kind of experiment. Thus Equation 10 represents a fundamental limitation on the measurement process itself!

Equation 10 is not like any formula we have previously dealt with in the text. It gives you an estimate, not an exact value. Often you will see the formula written  $\Delta p \Delta x \ge \hbar$  with  $\hbar = h/2\pi$ , rather than h, appearing on the right side. Whether you use h or  $\hbar$  depends upon how you wish to define the uncertainties  $\Delta p$  and  $\Delta x$ . But it is not necessary to be too precise. The important point is that the product  $\Delta p \Delta x$  must be at least of the order of magnitude h. It cannot be h/100 or something smaller.

The gist of the uncertainty principle is that the more accurately you measure the position of the particle, the more you mess up the particle's momentum. Or, the more accurately you measure the momentum of a particle, the less you know about the particle's position.

Equation (10) is not quite right, because it turns out that an accurate measurement of the x position of a particle does not necessarily mess up the particle's y component of momentum, only its x component. A more accurate statement of the uncertainty principle is

$$\Delta p_{x} \Delta x \ge h \tag{11a}$$

$$\Delta p_{\rm v} \Delta y \ge h \tag{11b}$$

where  $\Delta p_x$  is the uncertainty in the particle's x component of momentum due to a measurement of its x position, and  $\Delta p_y$  is the uncertainty in the y component of momentum resulting from a y position measurement. The quantities  $\Delta x$  and  $\Delta y$  are the uncertainty in the x and y measurements respectively.

## **Single Slit Experiment**

In our two slit thought experiment, we measured the position of the electron by hitting it with a photon. Another way to measure the position of a particle is to send it through a slit. For example, suppose a beam of particles impinges on a slit of width (w) as illustrated in Figure (11). We know that any particle that makes it to the far side of the slit had, at one time, been within the slit. At that time we knew its y position to within an uncertainty  $\Delta y$  equal to the width (w) of the slit.

$$\Delta y = w \tag{12}$$

This is an example of a position measurement with a precisely known uncertainty  $\Delta y$ .

According to the uncertainty principle, the particle's y component of momentum is uncertain by an amount  $\Delta p_y$  given by Equation 11b as

$$\Delta p_{y} \ge \frac{h}{\Delta y} = \frac{h}{w} \tag{13}$$

Equation 13 tells us that the smaller  $\Delta y$ , i.e., the narrower the slit, the bigger the uncertainty  $\Delta p_y$  we create in the particle's y momentum. This is what happens if the particle's motion is governed by its wave nature.

In Figure (12a) we have a ripple tank photograph of a wave passing through a moderately narrow slit. The wave on the far side of the slit is seen to spread out a bit. We can calculate the amount of spread by noting that the beam is mostly contained within the central maximum of the single slit diffraction pattern.





When a particle goes through the slit, its y position is known to within an uncertainty  $\Delta y = w$ .

Now suppose that this wave represented a beam of photons or electrons. On the right side of the slit, all the particles have a definite x component of momentum  $p_x = h/\lambda$  and no y momentum. The uncertainty in the y momentum is zero.

Once the particle's have gone through the slit, the beam spreads out giving the particles a y momentum.

Since you do not know whether any given particle in the beam will go straight ahead, up or down, the spread of the beam introduces an uncertainty  $\Delta p_y$  in the particle's y momentum. This spread is illustrated in Figure (12b).



**Figure 12a** *Wave picture of the single slit experiment, wide slit.* 



**Figure 12b** Particle picture of the single slit experiment.

In Figure (13a), we see a wave passing through a narrower slit than the one in Figure (12a). With the narrower slit, we have made a more precise measurement of the particle's y position. We have reduced the uncertainty  $\Delta y = w$ . According to the uncertainty principle  $\Delta p_y \ge h/\Delta y$ , a decrease in  $\Delta y$  should increase the uncertainty  $\Delta p_y$  in the particle's y momentum. But an increase in  $\Delta p_y$  means that the beam should spread out more, which is what it does in Figure (13). In going from Figure (12) to Figure (13), we have cut the slit width about in half and about doubled the spread. I.e., cutting  $\Delta y$  in half doubles  $\Delta p_y$  as expected.



**Figure 13a** Narrow the slit and the wave spreads out.



**Figure 13b** With a narrower slit,  $\Delta p_{y}$  increases.

#### Example 1

We can use our analysis of the single slit pattern in Chapter 33 to show that  $\Delta p_y$  and  $\Delta y$  are related by the uncertainty principle. We saw that when a wave goes through a single slit of width w, the distance  $y_{min}$  to the first minimum is given by

$$y_{\min} = \frac{\lambda D}{W}$$
(33-14)

where  $\lambda$  is the wavelength and D the distance to the screen as shown in Figure (14a). The angle to the first minimum is given by

$$\tan(\theta_{\min}) = \frac{y_{\min}}{D} = \frac{(\lambda D/w)}{D} = \frac{\lambda}{w}$$
(14)



#### Figure 14a

After the beam emerges from the slit, the momenta of the particles spreads out through the same angle  $\theta_{\text{min}}$  as indicated in Figure (14b). From that figure we have

$$\tan(\theta_{\min}) = \frac{\Delta p_{y}}{p_{x}}$$
(15)



#### Figure 14b

where  $\Delta p_y$  represents the possible spread in the particle's y momenta. Equating values of tan $(\theta_{min})$  from Equation 14 and Equation 15 gives

$$\frac{\lambda}{W} = \frac{\Delta \rho_{y}}{\rho_{x}}$$
(16)

The particles entered the slit as plane waves with only an x component of momentum given by de Broglie's formula

$$p_{x} = \frac{h}{\lambda}$$
(17)

Using this formula for  $p_x$  in Equation 16 gives

$$\frac{\lambda}{W} = \frac{\Delta p_y}{(h/\lambda)} = \frac{\lambda \Delta p_y}{h}$$
(18)

The  $\lambda$ 's cancel, and we are left with

$$\Delta p_{y} \bigg) w = h \tag{19}$$

But the slit width w is  $\Delta y$ , the uncertainty in the y measurement, thus

$$\Delta p_{v} \Delta y = h \tag{20}$$

There is an equal sign in Equation 20 because this particular measurement of the y position of the particle causes the least possible uncertainty in the particle's y component of momentum. (Note that the x component of the particle's momentum is more or less unaffected by the slit. The wave has the same wavelength  $\lambda$  before and after going through the slit. It is the y component of momentum that changed from zero on the left side to  $\pm \Delta p_{y}$  on the right.)

#### **Exercise 3**

A microwave beam, consisting of  $1.24 \times 10^{-4}$ eV photons impinges on a slit of width (w) as shown in Figure 14a.

(a) What is the momentum  $p_x$  of the photons in the laser beam before they get to the slit?

(b) When the photon passes through the slit, their y position is known to an uncertainty  $\Delta y = w$ , the slit width. Before the photons get to the slit, their y momentum has the definite value  $p_y = 0$ . Passing through the slit makes the photon's y momentum uncertain by an amount  $\Delta p_y$ . Using the uncertainty principle, calculate what the slit width (w) must be so that  $\Delta p_y$  is equal to the photon's original momentum  $p_x$ . How does w compare with the wavelength  $\lambda$  of the laser beam?

(c) If  $\Delta p_y$  becomes as large as the original momentum  $p_x$ , what can you say about the wave pattern on the right side of Figure (11)? Is this consistent with what you know about waves of wavelength  $\lambda$  passing through a slit of this width? Explain.

# TIME-ENERGY FORM OF THE UNCERTAINTY PRINCIPLE

The second form of the uncertainty principle, which perhaps has an even greater impact on our understanding of the behavior of matter, involves the measurement of the energy of a particle, and the time available to make the measurement. *The shorter the time available, the less accurate the energy measurement is.* If  $\Delta E$  is the uncertainty in the results of our energy measurement, and  $\Delta t$  the time we had to make the measurement, then  $\Delta E$  and  $\Delta t$  are related by

$$\Delta E \Delta t \ge h \tag{21}$$

One can derive this form of the uncertainty principle from  $\Delta p \Delta x \ge h$ , but we can gain a better insight into the relationship by starting with an explicit example.

A device that has become increasingly important in research, particularly in the study of fast reactions in molecules and atoms, is the pulsed laser. The lasers we have used in various experiments are all continuous beam lasers. The beam is at least as long as the distance from the laser to the wall. If we had a laser that we could turn on and off in one nanosecond, the pulse would be 1 foot or 30 cm long and contain

$$\frac{30 \text{ cm}}{6 \times 10^{-5} \frac{\text{ cm}}{\text{wavelength}}} = 5 \times 10^{5} \text{ wavelengths}$$

Even a picosecond laser pulse which is 1000 times shorter, contains 500 wavelengths. Some of the recent pulsed lasers can produce a pulse 500 times shorter than that, only 2 femtoseconds ( $2 \times 10^{-15}$  seconds) long. These lasers emit a pulse that is only one wavelength long. For our example of the time-energy form of the uncertainty principle, we wish to consider the nature of the photons in a 2 femtosecond long laser pulse. If we want to measure the energy of the photons in such a pulse, we only have 2 femtoseconds to make the measurement because that is how long the pulse takes to go by us. In the notation of the uncertainty principle

$$\Delta t = 2 \times 10^{-15} \text{sec} \qquad \begin{array}{l} \text{time available to} \\ \text{measure the energy} \\ \text{of the photons} \\ \text{in our laser pulse} \end{array}$$
(22)

Let us suppose that the laser produces red photons whose wavelength is  $6.2 \times 10^{-5}$  cm, about the wavelength of the lasers we have been using. According to our usual formula for calculating the energy of the photons in such a laser beam we have

$$E_{\text{photon}} = \frac{12.4 \times 10^{-5} \text{eV cm}}{6.2 \times 10^{-5} \text{cm}} = 2 \text{ eV}$$
(23)

Now let us use the uncertainty principle in the form

$$\Delta E \ge \frac{h}{\Delta t} \tag{24}$$

to calculate the uncertainty in any measurement we would make the energy of the photons in the 2 femtosecond laser beam. We have

$$\Delta E \ge \frac{h}{\Delta t} \ge \frac{6.63 \times 10^{-27} \text{erg sec}}{2 \times 10^{-15} \text{sec}}$$
$$\ge 3.31 \times 10^{-12} \text{ergs}$$
(25)

Converting  $\Delta E$  from ergs to electron volts, we get

$$\Delta E \ge \frac{3.31 \times 10^{-12} \text{ergs}}{1.6 \times 10^{-12} \text{ergs/eV}} \approx 2 \text{ eV}$$
(26)

The uncertainty in any energy measurement we make of these photons is as great as the energy itself! If we try to measure the energy of these photons, we expect the answers to range from  $E - \Delta E = 0 \text{ eV}$  up to  $E + \Delta E = 4 \text{ eV}$ . Why does this happen? Why is the energy of the photons in this beam so uncertain? *Fourier analysis provides the answer*. We can see why the energy of the photons in the 2 femtosecond pulse is so uncertain by comparing the Fourier transform of a long laser pulse with that of a pulse consisting of only one wavelength.

Figure (15) shows the Fourier transform of an infinitely long sine wave. You will recall that, in the design of the MacScope program, it is assumed that we are analyzing a repeated waveform. If you continuously repeat the waveform seen in the upper half of the diagram, you get an infinitely long cyclic wave which is a pure sine wave. (Sine waves are by definition infinitely long waves.) In effect we have in Figure (15) selected 16 cycles of the pure sine wave, and the Fourier analysis box shows that we have a pure 16th harmonic. This sine wave has a definite frequency f, and if this represented a laser beam, the photons in the beam would have a precise energy given by the formula E = hf. There is no uncertainty in the energy of this infinitely long sine wave. (It would take an infinite time  $\Delta t$  to make sure that the wave was infinitely long, with the result  $\Delta E = h/\Delta t = h/\infty = 0$ .)

 $\int_{-1}^{1} \frac{1}{\sqrt{1-1}} \frac{1}{$ 

**Figure 15** *A pure sine wave has a single frequency.* 

In Figure (16) we are looking at a waveform consisting of a single pulse. This would accurately represent the output of a red laser that continuously emitted single wavelength pulses spaced 16 wavelengths apart. (Remember that our program assumes that the wave shape is repeated.) From the Fourier analysis box we see that there is a dramatic difference between the composition of a pure sine wave and of a single pulse. To construct a single pulse out of sine waves, we have to add up a slew of harmonics. The single pulse is more like a drum beat while the continuous wave is more like a flute. (In the appendix we show how the sine wave harmonics add up to produce a pulse.)

In Figure (16) we see that the dominant harmonic is still around the sixteenth, as it was for the continuous wave, but there is a spread of harmonics from near zero up to almost the 32nd. For a laser pulse to have this shape, it must consist of frequencies ranging from near zero up to twice the natural frequency. Each of these frequencies contains photons whose energy is given by Einstein's formula E = hf where f is the frequency of the harmonic.



One cycle of a wave is made up of a spread of harmonics

In Figure (17) we have reproduced the Fourier analysis box of Figure (16), but relabeled the horizontal axis in electron volts. We have assumed that the 16th harmonic represented 2 eV photons which would be the case if the wave were an infinitely long red laser pulse. Now the diagram represents the density of photons of different energies in the laser beam. While 2 eV is the most likely energy, there is a spread of energies ranging from nearly 0 eV up to nearly 4 eV. If we measure the energy of a photon in the beam, our answer is 2 eV with an uncertainty of 2 eV, just as predicted by the uncertainty principle  $\Delta E \Delta t \ge h$ .



**Figure 17** *Photon energies in single wavelength pulse of a red laser beam.* 



**Figure 18** A two cycle wave has half the spread of harmonics.

In Figure (18) we analyze a pulse two wavelengths long. Now we see that the spread of frequencies required to reconstruct this waveform is only half as wide, ranging from the 8th to 24th harmonic, or from 1 eV to 3 eV. We have twice as long to study a 2 wavelength pulse, and the uncertainty in energy  $\Delta E$  is only about 1 eV, or half as big.

Going to a 4 wavelength pulse in Figure (19) we see that by doubling the time available we again cut in half the uncertainty  $\Delta E$  in energy. Now the energy varies from about 1.5 eV to 2.5 eV for a  $\Delta E = .5$  eV. This is just what you expect from  $\Delta E \Delta t \ge h$ . You should now begin to see that the uncertainty principle is a simple rule evolving from the wave nature of particles. (By the way, it would be more accurate to write  $\Delta E \Delta t = h$  for this discussion, because we are describing the very least uncertainty in energy.)



**Figure 19** A four cycle wave has a fourth the spread of harmonics.

# **Probability Interpretation**

We have interpreted Figure (17) as representing the spread in energies of the photons in a 2 femtosecond red laser pulse. What if the pulse consisted of only a single photon? Then how do we interpret this spread in energies? The answer is that we use a probability interpretation. The photon in the pulse has different probabilities of having different energies.

In our discussion of light waves, we saw that the energy density in a light wave was proportional to the square of the amplitude of the wave. This is reasonable because while the amplitude of a wave can be positive or negative, the square of the amplitude, which we call the *intensity* is always positive. Probabilities, like energy densities, also have to be positive, thus we should associate the probability of a photon as having a given frequency with the intensity or square of the amplitude of the wave of that frequency.

In Figure (20) we show the intensities (square of the amplitudes) of the harmonics that make up the single wavelength pulse. (This is plotted automatically by MacScope when we click on the button labeled  $\Phi$ .) We see that squaring the amplitudes narrows the spread.

Figure (20) has the following interpretation when applied to pulses containing a single photon. If we measure the energy of the photon, we are most likely to get an answer close to 2 eV but there is a reasonable probability of getting an answer lower than 1 eV or even higher than 3 eV. The heights of the bars tell us the relative probability of measuring that energy for the photon.



#### Figure 20

Intensities of the harmonics are proportional to the square of the amplitudes.

# **Measuring Short Times**

We have said that the new pulsed lasers produce pulses as short as 2 femtoseconds. How do we know that? Suppose we gave you the job of measuring the length of the laser pulse, and the best oscilloscope you had could measure times no shorter than a nanosecond. This is a million times too slow to see a femtosecond pulse. What do you do?

If you cannot measure the time directly, you can be sneaky and use the uncertainty principle. Send the laser pulse through a diffraction grating, and record the spread in wavelengths, i.e., the spread in energies of the photons in the pulse. If the line is very sharp, if they are all red photons of a single wavelength and energy, then you know that there is no measurable uncertainty  $\Delta E$ in the photon energies, and the pulse must last a time  $\Delta t$  that is considerably longer than 2 femtoseconds. If, on the other hand, the line is spread out from the near infra red to violet, if the spread in energies is from 1 eV to 3 eV, and the spread is not caused by some other phenomena (like the Doppler effect), then from the uncertainty principle you know that the pulse is only about a femtosecond long. (You know, for example, it cannot be as long as 10 femtoseconds, or as short as a tenth of a femtosecond.)

Thus, with the uncertainty principle, you can use a diffraction grating rather than a clock or oscilloscope to measure very short times. Instead of being an annoying restriction on our ability to make experimental measurements, the uncertainty principle can be turned into an important scientific tool for measuring short times and, as we shall see, short distances.

## **Exercise 4**

An electron is in an excited state of the hydrogen atom, either the second energy level at -3.40 eV or the third energy level at -1.51 eV. You wish to do an experiment to decide which of these two states the electron is in. What is the least amount of time you **must** take to make this measurement?

## **Short Lived Elementary Particles**

We usually think of the rest energy of a particle as having a definite value. For example the rest energy of a proton is  $938.2723 \times 10^6 \text{eV}$ . The proton itself is a composite particle made of 3 quarks, and the number 938.2723 MeV represents the total energy of the quarks in the allowed wave pattern that represents a proton. This rest energy has a very definite value because the proton is a stable particle with plenty of time to settle into a precise wave pattern.

A rather different particle is the so called " $\Lambda(1520)$ ", which is another combination of 3 quarks, but very short lived. The name comes partly from the fact that the particle's rest mass energy is about 1520 million electron volts (MeV). As indicated in Figure (21), a  $\Lambda(1520)$  can be created as a result of the collision between a K<sup>-</sup> meson and a proton.

We are viewing the collision in a special coordinate system, where the total momentum of the incoming particles is zero. In this coordinate system, the resulting  $\Lambda(1520)$  will be at rest. By conservation of energy, the total energy of the incoming particles should equal the



a) A K<sup>-</sup>meson and a proton are about to collide. We are looking at the collision in a coordinate system where the total momentum is zero (the so called "center of mass" system).

b) In the collision a  $\Lambda(1520)$  particle is created. It is at rest in this center of mass system



c) The  $\Lambda(1520)$  then quickly decays into a lower energy  $\Lambda$  particle and two  $\pi$  mesons.

#### Figure 21

A  $\Lambda(1520)$  particle can be created if the total energy (in the center of mass system) of the incoming particles equals the rest mass energy of the  $\Lambda(1520)$ . rest mass energy of the  $\Lambda$  (1520). Thus if we collide K<sup>-</sup> particles with protons, we expect to create a  $\Lambda$  (1520) particle only if the incoming particles have the right total energy.

Figure (22) shows the results of some collision experiments, where a K<sup>-</sup> meson and a proton collided to produce a  $\Lambda$  and two  $\pi$  mesons. The probability of such a result peaked when the energy of the incoming particles was 1,520 MeV. This peak occurred because the incoming K<sup>-</sup> meson and proton created a  $\Lambda$  (1520) particle, which then decayed into the  $\Lambda$  and two  $\pi$  mesons, as shown in Figure (21). The  $\Lambda$  (1520) was not observed directly, because its lifetime is too short.

Figure (22) shows that the energy of the incoming particles does not have to be exactly 1520 MeV in order to create a  $\Lambda$  (1520). The peak is in the range from about 1510 to 1530 MeV, which implies that the rest mass energy of the  $\Lambda$  (1520) is 1520 MeV plus or minus about 10 MeV. From one experiment to another, the rest mass energy can vary by about 20 MeV. (The experimentalists quoted a variation of 16 MeV.)



#### Figure 22

The probability that a K<sup>-</sup>meson and a proton collide to produce a  $\Lambda$  particle, and two  $\pi$  mesons peak at an energy of 1520 MeV. The peak results from the fact that a  $\Lambda(1520)$  particle was created and quickly decayed into the  $\Lambda$  and two  $\pi$  mesons. The probability peaks at 1520 MeV, but can be seen to spread out over a range of about 16MeV. The small circles are experimental values, the vertical lines represent the possible error in the value. (Data from M.B. Watson et al., Phys. Rev. 131(1963).) Why isn't the peak sharp? Why does the rest mass energy of the  $\Lambda$  (1520) particle vary by as much as 16 to 20 MeV from one experiment to another? The answer lies in the fact that *the lifetime of the*  $\Lambda$ (1520) *is so short, that the particle does not have enough time to establish a definite rest mass energy*. The 16 MeV variation is the uncertainty  $\Delta E$  in the particles rest mass energy that results from the fact that the particle's lifetime is limited.

The uncertainty principle relates the uncertainty in energy  $\Delta E$  to the time  $\Delta t$  available to establish that energy. To establish the rest mass energy, time  $\Delta t$ available is the particle's *lifetime*. Thus we can use the uncertainty principle to estimate the lifetime of the  $\Lambda$  (1520) particle. With  $\Delta E \times \Delta t \approx h$  we get

$$\Delta t \approx \frac{h}{\Delta E} = \frac{6.63 \times 10^{-27} \text{erg sec}}{16 \text{ MeV} \times \left(1.6 \times 10^{-6} \frac{\text{erg}}{\text{MeV}}\right)}$$
$$\Delta t \approx 2.6 \times 10^{-22} \text{ seconds}$$
(27)

The lifetime of the  $\Lambda$  (1520) particle is of the order of  $10^{-22}$  seconds! This is only about 10 times longer than it takes light to cross a proton! Only by using the uncertainty principle could we possibly measure such short times.

# THE UNCERTAINTY PRINCIPLE AND ENERGY CONSERVATION

The fact that for short times the energy of a particle is uncertain, raises an interesting question about basic physical laws like the law of conservation of energy. If a particle's energy is uncertain, how do we know that energy is conserved in some process involving that particle? The answer is -- *we don't*.

One way to explain the situation is to say that nature will cheat if it can get away with it. Energy does not have to be conserved if we cannot do an experiment to demonstrate a lack of conservation of energy.

Consider the process shown in Figure (23). It shows a red, 2 eV photon traveling along in space. Suddenly the photon creates an positron-electron pair. The rest mass energy of both the positron and the electron are .51 MeV. Thus we have a 2 eV photon creating a pair of particles whose total energy is  $1.02 \times 10^6$ eV, a huge violation of the law of conservation of energy. A short time later the electron and positron come back together, annihilate, leaving behind a 2 eV photon. This is an equally huge violation of the conservation of energy.

But have we really violated the conservation of energy? During its lifetime, the positron-electron pair is a composite object whose total energy is uncertain. If the pair lived a long time, its total energy would be close to the expected energy of  $1.02 \times 10^6 \text{eV}$ . But suppose the pair were in existence only for a very short time  $\Delta t$ , a time so short that the uncertainty in the energy could be as large as  $1.02 \times 10^6 \text{eV}$ . Then there is some probability that the energy of the pair might be only 2 eV and the process shown in Figure (2) could happen.



**Figure 23** Consider a process where a 2 eV photon suddenly creates a positron-electron pair. A short time later the pair annihilates, leaving a 2 eV photon. In the long range, energy is conserved.

The length of time  $\Delta t$  that the pair could exist and have an energy uncertain by 1.02 MeV is

$$\Delta t = \frac{h}{\Delta E} = \frac{6.63 \times 10^{-27} \text{ erg sec}}{1.02 \text{ MeV} \times 1.6 \times 10^{-6} \frac{\text{erg}}{\text{MeV}}}$$
$$\Delta t = 4 \times 10^{-21} \text{sec}$$
(32)

Another way to view the situation is as follows. Suppose the pair in Figure (20) lasted only  $4 \times 10^{-21}$  seconds or less. Even if the pair had an energy of  $1.02 \times 10^{6} \text{eV}$ , the lifetime is so short that any measurement of the energy of the pair would be uncertain by at least  $1.02 \times 10^{6} \text{eV}$ , and the experiment could not detect the violation of the law of conservation of energy. In this point of view, if we cannot perform an experiment to detect a violation of the conservation law, then the process should have some probability of occurring.

Does a process like that shown in Figure (23) actually occur? If so, is there any way that we can know that it does? The answer is yes, to both questions. It is possible to make extremely accurate studies of the energy levels of the electron in hydrogen, and to make equally accurate predictions of the energy using the theory of *quantum electrodynamics*. We can view the binding of the electron in hydrogen as resulting from the continual exchange of photons between the electron and proton. During this continual exchange, there is some probability that the photon creates a positron electron pair that quickly annihilates as shown in Figure (23). In order to predict the correct values of the hydrogen energy levels, the process shown in Figure (23) has to be included. Thus we have direct experimental evidence that for a short time the particle antiparticle pair existed.

# QUANTUM FLUCTUATIONS AND EMPTY SPACE

We began the text with a discussion of the principle of relativity—that you could not detect your own motion relative to empty space. The concept of empty space seemed rather obvious—space with nothing in it. But the idea of empty space is not so obvious after all.

With the discovery of the cosmic background radiation, we find that all the space in this universe is filled with a sea of photons left over from the big bang. We can accurately measure our motion relative to this sea of photons. The earth is moving relative to this sea at a velocity of 600 kilometers per second toward the Vergo cluster of galaxies. While this measurement does not violate the principle of relativity, it is in some sense a measurement of our motion relative to the universe as a whole.

Empty space itself may not be empty. Consider a process like that shown in Figure (24) where a photon, an electron, and a positron are all created at some point in space. A short while later the three particles come back together with the positron and electron annihilating and the photon being absorbed.

One's first reaction might be that such a process is ridiculous. How could these three particles just appear and then disappear? To do this we would have to violate both the laws of conservation of energy and momentum.

But, of course, the uncertainty principle allows us to do that. We can, in fact, use the uncertainty principle to estimate how long such an object could last. The arguments would be similar to the ones we used in the analysis of the process shown in Figure (23).



**Figure 24** *Quantum fluctuation. The uncertainty principle allows such an object to suddenly appear, and then disappear.* 

In the theory of quantum electrodynamics, a completely isolated process like that shown in Figure (24) does not affect the energy levels of the hydrogen atom and should be undetectable in electrical measurements. But such a process might affect gravity. A gravitational wave or a graviton might interact with the energy of such an object. Some calculations have suggested that such interactions could show up in Einstein's classical theory of gravity as a correction to the famous *cosmological constant* we discussed in Chapter 21.

An object like that shown in Figure (23) is an example of what one calls a *quantum fluctuation*. Here we have something that appears and disappears in so-called empty space. If such objects can keep appearing and disappearing, then we have to revise our understanding of what we mean by empty.

The uncertainty principle allows us to tell the difference between a quantum fluctuation and a real particle. A quantum fluctuation like that in Figure (24) violates conservation of energy, and therefore cannot last very long. A real particle can last a long time because energy conservation is not violated.

However, there is not necessarily that much difference between a real object and a quantum fluctuation. To see why, let us take a closer look at the  $\pi$  meson. The  $\pi^+$ is a particle with a rest mass energy of 140 MeV, that consists of a quark-antiquark pair. The quark in that pair is the so-called **up** quark that has a rest mass of roughly 400 MeV. The other is the **antidown** quark that has a rest mass of about 700 MeV. (Since we can't get at isolated quarks, the quark rest masses are estimates, but should not be too far off). Thus the two quarks making up the  $\pi$  meson have a total rest mass of about 1100 MeV. How could they combine to produce a particle whose rest mass is only 140 MeV? The answer lies in the potential energy of the gluon force that holds the quarks together. As we have seen many times, the potential energy of an attractive force is negative. In this case the potential energy of the gluon force is almost as big in magnitude as the rest mass of the quarks, reducing the total energy from 1100 MeV to 140 MeV.

Suppose we had an object whose negative potential energy was as large as the positive rest mass energy. Imagine, for example, that the object consisted of a collection of point sized elementary particles so close together that their negative gravitational potential energy was the same magnitude as the positive rest mass and kinetic energy. Suppose such a collection of particles were created in a quantum fluctuation. How long could the fluctuation last?

Since such an object has no total energy, the violation  $\Delta E$  of energy conservation is zero, and therefore the lifetime  $\Delta t = h/\Delta E$  could be forever.

Suppose the laws of physics required that such a fluctuation rapidly expand, greatly increasing both the positive rest mass and kinetic energy, while maintaining the corresponding amount of negative gravitational potential energy. As long as  $\Delta E$  remained zero, the expanding fluctuation could keep on going. Perhaps such a fluctuation occurred 14 billion years ago and we live in it now.

# **APPENDIX**

# HOW A PULSE IS FORMED FROM SINE WAVES



**Figure A1** By selecting more and more harmonics, you can

see how the sine waves add up to produce a pulse.

# Chapter on **Geometrical Optics**

For over 100 years, from the time of Newton and Huygens in the late 1600s, until 1801 when Thomas Young demonstrated the wave nature of light with his two slit experiment, it was not clear whether light consisted of beams of particles as proposed by Newton, or was a wave phenomenon as put forward by Huygens. The reason for the confusion is that almost all common optical phenomena can be explained by tracing light rays. The wavelength of light is so short compared to the size of most objects we are familiar with, that light rays produce sharp shadows and interference and diffraction effects are negligible.





*To see how wave phenomena can be explained by ray* tracing, consider the reflection of a light wave by a metal surface. When a wave strikes a very small object, an object much smaller than a wavelength, a circular scattered wave emerges as shown in the ripple tank photograph of Figure (36-1) reproduced here. But when a light wave impinges on a metal surface consisting of many small atoms, represented by the line of dots in Figure (36-2), the circular scattered waves all add up to produce a reflected wave that emerges at an angle of reflection  $\theta_r$  equal to the angle of incidence  $\theta_i$ . Rather than sketching the individual crests and troughs of the incident wave, and adding up all the scattered waves, it is much easier to treat the light as a ray that reflected from the surface. This ray is governed by the law of reflection, namely  $\theta_r = \theta_i$ .



#### Figure 36-2

Reflection of light. In the photograph, we see an incoming plane wave scattered by a small object. If the object is smaller than a wavelength, the scattered waves are circular. When an incoming light wave strikes an array of atoms in the surface of a metal, the scattered waves add up to produce a reflected wave that comes out at an angle of reflection  $\theta_r$  equal to the angle of incidence  $\theta_i$ .

The subject of geometrical optics is the study of the behavior of light when the phenomena can be explained by ray tracing, where shadows are sharp and interference and diffraction effects can be neglected. The basic laws for ray tracing are extremely simple. At a reflecting surface  $\theta_r = \theta_i$ , as we have just seen. When a light ray passes between two media of different **indexes of refraction**, as in going from air into glass or air into water, the rule is  $n_1 \sin \theta_1 = n_2 \sin \theta_2$ , where  $n_1$  and  $n_2$  are constants called indices of refraction, and  $\theta_1$  and  $\theta_2$  are the angles that the rays made with the line perpendicular to the interface. This is known as **Snell's law**.

This entire chapter is based on the two rules  $\theta_r = \theta_i$ and  $n_1 \sin \theta_1 = n_2 \sin \theta_2$ . These rules are all that are needed to understand the function of telescopes, microscopes, cameras, fiber optics, and the optical components of the human eye. You can understand the operation of these instruments without knowing anything about Newton's laws, kinetic and potential energy, electric or magnetic fields, or the particle and wave nature of matter. In other words, there is no prerequisite background needed for studying geometrical optics as long as you accept the two rules which are easily verified by experiment. In most introductory texts, geometrical optics appears after Maxwell's equations and theory of light. There is a certain logic to this, first introducing a basic theory for light and then treating geometrical optics as a practical application of the theory. But this is clearly not an historical approach since geometrical optics was developed centuries before Maxwell's theory. Nor is it the only logical approach, because studying lens systems teaches you nothing more about Maxwell's equations than you can learn by deriving Snell's law. Geometrical optics is an interesting subject full of wonderful applications, a subject that can appear anywhere in an introductory physics course.

We have a preference not to introduce geometrical optics after Maxwell's equations. With Maxwell's theory, the student is introduced to the wave nature of one component of matter, namely light. If the focus is kept on the basic nature of matter, the next step is to look at the photoelectric effect and the particle nature of light. You then see that light has both a particle and a wave nature, which opens the door to the particle-wave nature of all matter and the subject of quantum mechanics. We have a strong preference not to interrupt this focus on the basic nature of matter with a long and possibly distracting chapter on geometrical optics.

## **REFLECTION FROM CURVED SURFACES**

The Mormon Tabernacle, shown in Figure (1), is constructed in the shape of an ellipse. If one stands at one of the focuses and drops a pin, the pin drop can be heard 120 feet away at the other focus. The reason why can be seen from Figure (2), which is similar to Figure (8-28) where we showed you how to draw an ellipse with a pencil, a piece of string, and two thumbtacks.

The thumbtacks are at the focuses, and the ellipse is drawn by holding the string taut as shown. As you move the pencil point along, the two sections of string always make equal angles  $\theta_i$  and  $\theta_r$  to a line perpen-



Mormon Tabernacle under construction, 1866.



Mormon Tabernacle finished, 1871.



*Mormon Tabernacle today.* **Figure 1** 

dicular or normal to the part of the ellipse we are drawing. The best way to see that the angles  $\theta_i$  and  $\theta_r$  are always equal is to construct your own ellipse and measure these angles at various points along the curve.

If a sound wave were emitted from focus 1 in Figure (2), the part of the wave that traveled over to point A on the ellipse would be reflected at an angle  $\theta_r$  equal to the angle of incidence  $\theta_i$ , and travel over to focus 2. The part of the sound wave that struck point B on the ellipse, would be reflected at an angle  $\theta_r$  equal to it's angle of incidence  $\theta_i$ , and also travel over to focus 2. If you think of the sound wave as traveling out in rays, then all the rays radiated from focus 1 end up at focus 2, and that is why you hear the whisper there. We say that the rays are *focused* at focus 2, and that is why these points are called focuses of the ellipse. (Note also that the path lengths are the same, so that all the waves arriving at focus 2 are in phase.)



**Figure 2** Drawing an ellipse using a string and two thumbtacks.



**Figure 2a** A superposition of the top half of Figure 2 on Figure 1.

# **The Parabolic Reflection**

You make a parabola out of an ellipse by moving one of the focuses very far away. The progression from a parabola to an ellipse is shown in Figure (3). For a true parabola, the second focus has to be infinitely far away.

Suppose a light wave were emitted from a star and traveled to a parabolic reflecting surface. We can think of the star as being out at the second, infinitely distant, focus of the parabola. Thus all the light rays coming in from the star would reflect from the parabolic surface and come to a point at the near focus. The rays from the star approach the reflector as a parallel beam of rays, thus a parabolic reflector has the property of focusing parallel rays to a point, as shown in Figure (4a).





If parallel rays enter a deep dish parabolic mirror from an angle off axis as shown in Figure (4b), the rays do not focus to a point, with the result that an off axis star would appear as a blurry blob. (This figure corresponds to looking at a star  $2.5^{\circ}$  off axis, about 5 moon diameters from the center of the field of view.)



**Figure 4b** For such a deep dish par

For such a deep dish parabola, rays coming in at an angle of  $2.5^{\circ}$  do not focus well.

f2 mirror

One way to get sharp images for parallel rays coming in at an angle is to use a shallower parabola as illustrated in Figure (4c). In that figure, the *focal length* (distance from the center of the mirror to the focus) is 2 times the mirror diameter, giving what is called an f 2 mirror. In Figure (4d), you can see that rays coming in at an angle of  $2.5^{\circ}$  (blue lines) almost focus to a point. Typical amateur telescopes are still shallower, around f8, which gives a sharp focus for rays off angle by as much as  $2^{\circ}$  to  $3^{\circ}$ .

As we can see in Figure (4d), light coming from two different stars focus at two different points in what is called the *focal plane* of the mirror. If you placed a photographic film at the focal plane, light from each different star, entering as parallel beams from different angles, would focus at different points on the film, and you would end up with a photographic image of the stars. This is how distant objects like stars are photographed with what is called a *reflecting telescope*.

## light from star on axis



#### Figure 4c

A shallow dish is made by using only the shallow bottom of the parabola. Here the focal length is twice the diameter of the dish, giving us an f2 mirror. Typical amateur telescopes are still shallower, having a focal length around 8 times the mirror diameter (f8 mirrors). [The mirror in Figure 4b, that gave a bad focus, was f.125, having a focal length 1/8 the diameter of the mirror.]



light from star #1, on axis

#### Figure 4d

We can think of this drawing as representing light coming in from a red star at the center of the field of view, and a blue star 2.5° (5 full moon diameters) away. Separate images are formed, which could be recorded on a photographic film. With this shallow dish, the off axis image is sharp (but not quite a point).

# **MIRROR IMAGES**

The image you see in a mirror, although very familiar, is still quite remarkable in its reality. Why does it look so real? You do not need to know how your eye works to begin to see why.

Consider Figure (5a) where light from a point source reaches your eye. We have drawn two rays, one from the source to the top of the eye, and one to the bottom. In Figure (5b), we have placed a horizontal mirror as shown and moved the light source a distance h above the mirror equal to the distance it was below the mirror before the mirror was inserted. Using the rule that the angle of incidence equals the angle of reflection, we again drew two rays that went from the light source to



Figure 5a

Light from a point source reaching your eye.



Figure 5b

There is no difference when the source is at point A, or at point A' and the light is reflected in a mirror.

the top and to the bottom of the eye. You can see that if you started at the eye and drew the rays back as straight lines, ignoring the mirror, the rays would intersect at the old source point A as shown by the dotted lines in Figure (5b).

To the eye (or a camera) at point B, there is no detectable difference between Figures (5a) and (5b). In both cases, the same rays of light, coming from the same directions enter the eye. Since the eye has no way of telling that the rays have been bent, we perceive that the light source is at the *image point* A rather than at the source point A'.

When we look at an extended object, its image in the mirror does not look identical to the object itself. In Figure (6), my granddaughter Julia is holding her right hand in front of a mirror and her left hand off to the side. The image of the right hand looks like the left hand. In particular, the fingers of the mirror image of the right hand curl in the opposite direction from those of the right hand itself. If she were using the right hand rule to find the direction of the angular momentum of a rotating object, the mirror image would look as if she were using a left hand rule.

It is fairly common knowledge that left and right are reversed in a mirror image. But if left and right are reversed, why aren't top and bottom reversed also? Think about that for a minute before you go on to the next paragraph.



**Figure 6** *The image of the right hand looks like a left hand.* 

To see what the image of an extended object should be, imagine that we place an arrow in front of a mirror as shown in Figure (7). We have constructed rays from the tip and the base of the arrow that reflect and enter the eye as shown. Extending these rays back to the image, we see that the image arrow has been reversed *front to back*. That is what a mirror does. The mirror image is reversed front to back, not left to right or top to bottom. It turns out that the right hand, when reversed front to back as in its image in Figure (6), has the symmetry properties of a left hand. If used to define angular momentum, you would get a left hand rule.



**Figure 7** *A mirror image changes front to back, not left to right.* 



Figure 8a

With a corner reflector, the light is reflected back it the same direction from which it arrived.

## The Corner Reflector

When two vertical mirrors are placed at right angles as shown in Figure (8a), a horizontal ray approaching the mirrors is reflected back in the direction from which it came. It is a little exercise in trigonometry to see that this is so. Since the angle of incidence equals the angle of reflection at each mirror surface, we see that the angles labeled  $\theta_1$  must be equal to each other and the same for the angles  $\theta_2$ . From the right triangle ABC, we see that  $\theta_1 + \theta_2 = 90^\circ$ . We also see that the angles  $\theta_2 + \theta_3$  also add up to 90°, thus  $\theta_3 = \theta_1$ , which implies the exiting ray is parallel to the entering one.

If you mount three mirrors perpendicular to each other to form the corner of a cube, then light entering this so called *corner reflector* from any angle goes back in the direction from which it came. The Apollo II astronauts placed the array of corner reflectors shown in Figure (8b) on the surface of the moon, so that a laser beam from the earth would be reflected back from a precisely known point on the surface of the moon. By measuring the time it took a laser pulse to be reflected back from the array, the distance to the moon could be measured to an accuracy of centimeters. With the distance to the moon known with such precision, other distances in the solar system could then be determined accurately.



#### Figure 8b

Array of corner reflectors left on the moon by the Apollo astronauts. A laser pulse from the earth, aimed at the reflectors, returns straight back to the laser. By measuring the time the pulse takes to go to the reflectors and back, the distance to that point on the moon and back can be accurately measured.

## **MOTION OF LIGHT THROUGH A MEDIUM**

We are all familiar with the fact that light can travel through clear water or clear glass. With some of the new glasses developed for fiber optics communication, light signals can travel for miles without serious distortion. If you made a mile thick pane from this glass you could see objects through it.

From an atomic point of view, it is perhaps surprising that light can travel any distance at all through water or glass. A reasonable picture of what happens when a light wave passes over an atom is provided by the ripple tank photograph shown in Figure (36-1) reproduced here. The wave scatters from the atom, and since atoms are considerably smaller than a wavelength of visible



a) Incident and scattered wave together.

light, the scattered waves are circular like those in the ripple tank photograph. The final wave is the sum of the incident and the scattered waves as shown in Figure (36-1a).

When light passes through a medium like glass or water, the wave is being scattered by a huge number of atoms. The final wave pattern is the sum of the incident wave and all of the many billions of scattered waves. You might suspect that this sum would be very complex, but that is not the case. At the surface some of the incident wave is reflected. Inside the medium, the *incident and scattered waves add up to a new wave* of the same frequency as the incident wave but which travels *at a reduced speed*. The speed of a light wave in water for example is 25% less than the speed of light in a vacuum.



b) After incident wave has passed.

#### Figure 36-1

If the scattering object is smaller than a wavelength, we get circular scattered waves.

The optical properties of lenses are a consequence of this effective reduction in the speed of light in the lens. Figure (9) is a rather remarkable photograph of individual short pulses of laser light as they pass through and around a glass lens. You can see that the part of the wave front that passed through the lens is delayed by its motion through the glass. The thicker the glass, the greater the delay. You can also see that the delay changed the shape and direction of motion of the wave front, so that the light passing through the lens focuses to a point behind the lens. This is how a lens really works.



#### Figure 9

Motion of a wave front through a glass lens. The delay in the motion of the wave front as it passes through the glass changes the shape and direction of motion of the wave front, resulting in the focusing of light. (This photograph should not be confused with ripple tank photographs where wavelengths are comparable to the size of the objects. Here the wavelength of the light is about one hundred thousand times smaller than the diameter of the lens, with the result we get sharp shadows and do not see diffraction effects.)

In the 18/February/1999 issue of Nature it was announced that a laser pulse travelled through a gas of supercooled sodium atoms at a speed of 17 meters per second! (You can ride a bicycle faster than that.) This means that the sodium atoms had an index of refraction of about 18 million, 7.3 million times greater than that of diamond!

## Index of Refraction

The amount by which the effective speed of light is reduced as the light passes through a medium depends both upon the medium and the wavelength of the light. There is very little slowing of the speed of light in air, about a 25% reduction in speed in water, and nearly a 59% reduction in speed in diamond. In general, blue light travels somewhat slower than red light in nearly all media.

It is traditional to describe the slowing of the speed of light in terms of what is called the *index of refraction* of the medium. The index of refraction n is defined by the equation

speed of light  
in a medium 
$$v_{light} = \frac{c}{n}$$
 (1)

The index n has to equal 1 in a vacuum because light always travels at the speed  $3 \times 10^8$  meters in a vacuum. The index n can never be less than 1, because nothing can travel faster than the speed c. For yellow sodium light of wavelength  $\lambda = 5.89 \times 10^{-5}$  cm (589 nanometers), the index of refraction of water at 20° C is n = 1.333, which implies a 25% reduction in speed. For diamond, n = 2.417 for this yellow light. Table 1 gives the indices of refraction for various transparent substances for the sodium light.

1.00000	exactly
1.00029	
1.309	
1.333	
1.36	
1.46	
1.49	
1.52	
1.54	
1.55	
1.65	
1.77	
1.923	
2.417	
2.907	
3.50	
18000000	for laser pulse
	$\begin{array}{c} 1.00000\\ 1.00029\\ 1.309\\ 1.333\\ 1.36\\ 1.46\\ 1.49\\ 1.52\\ 1.54\\ 1.55\\ 1.65\\ 1.77\\ 1.923\\ 2.417\\ 2.907\\ 3.50\\ 18000000\end{array}$

#### Table 1

Some indices of refraction for yellow sodium light at a wavelength of 589 nanometers.

## **Exercise 1a**

What is the speed of light in air, water, crown glass, and diamond. Express your answer in feet/nanosecond. (Take c to be exactly 1 ft/nanosecond.)

## **Exercise 1b**

In one of the experiments announced in *Nature*, a laser pulse took 7.05 microseconds to travel .229 millimeters through the gas of supercooled sodium atoms. What was the index of refraction of the gas for this particular experiment? (The index quoted on the previous page was for the slowest observed pulse. The pulse we are now considering went a bit faster.)

# **CERENKOV RADIATION**

In our discussion, in Chapter 1, of the motion of light through empty space, we saw that nothing, not even information, could travel faster than the speed of light. If it did, we could, for example, get answers to questions that had not yet been thought of.

When moving through a medium, the speed of a light wave is slowed by repeated scattering and it is no longer true that nothing can move faster than the speed of light in that medium. We saw for example that the speed of light in water is only 3/4 the speed c in vacuum. Many elementary particles, like the muons in the muon lifetime experiment, travel at speeds much closer to c. When a charged particle moves faster than the speed of light in a medium, we get an effect not unlike the sonic boom produced by a supersonic jet. We get a *shock* wave of light that is similar to a sound shock wave (sonic boom), or to the water shock wave shown in Figure (33-30) reproduced here. The light shock wave is called Cerenkov radiation after the Russian physicist Pavel Cerenkov who received the 1958 Nobel prize for discovering the effect.

In the muon lifetime picture, one observed how long muons lived when stopped in a block of plastic. The experiment was made possible by Cerenkov radiation. The muons that stopped in the plastic, entered moving faster than the speed of light in plastic, and as a result emitted a flash of light in the form of Cerenkov radiation. When the muon decayed, a charged positron and a neutral neutrino were emitted. In most cases the charged positron emerged faster than the speed of light in the plastic, and also emitted Cerenkov radiation. The two flashes of light were detected by the phototube which converted the light flashes to voltage pulses. The voltage pulses were then displayed on an oscilloscope screen where the time interval between the pulses could be measured. This interval represented the time that the muon lived, mostly at rest, in the plastic.



#### Figure 33-30

When the source of the waves moves faster than the speed of the waves, the wave fronts pile up to produce a shock wave as shown. This shock wave is the sonic boom you hear when a jet plane flies overhead faster than the speed of sound.

## SNELL'S LAW

When a wave enters a medium of higher index of refraction and travels more slowly, the wavelength of the wave changes. The wavelength is the distance the wave travels in one period, and if the speed of the wave is reduced, the distance the wave travels in one period is reduced. (In most cases, the frequency or period of the wave is not changed. The exceptions are in fluorescence and nonlinear optics where the frequency or color of light can change.)

We can calculate how the wavelength changes with wave speed from the relationship

$$\lambda \frac{\mathrm{cm}}{\mathrm{cycle}} = \frac{\mathrm{v_{wave}} \frac{\mathrm{cm}}{\mathrm{sec}}}{\mathrm{T} \frac{\mathrm{sec}}{\mathrm{cycle}}}$$

Setting  $v_{wave} = c/n$  for the speed of light in the medium, gives for the corresponding wavelength  $\lambda_n$ 

$$\lambda_{n} = \frac{v_{wave}}{T} = \frac{c/n}{T} = \frac{1}{n}\frac{c}{T} = \frac{\lambda_{0}}{n}$$
(2)

where  $\lambda_0 = c/T$  is the wavelength in a vacuum. Thus, for example, the wavelength of light entering a diamond from air will be shortened by a factor of 1/2.42.

What happens when a set of periodic plane waves goes from one medium to another is illustrated in the ripple tank photograph of Figure (10). In this photograph, the



#### Figure 10

Refraction at surface of water. When the waves enter shallower water, they travel more slowly and have a shorter wavelength. The waves must travel in a different direction in order for the crests to match up. water has two depths, deeper on the upper part where the waves travel faster, and shallower in the lower part where the waves travel more slowly. You can see that the wavelengths are shorter in the lower part, but there are the same number of waves. (We do not gain or loose waves at the boundary.) The frequency, the number of waves that pass you per second, is the same on the top and bottom.

The only way that the wavelength can be shorter and still have the same number of waves is for the wave to bend at the boundary as shown. We have drawn arrows showing the direction of the wave in the deep water (the incident wave) and in the shallow water (what we will call the *transmitted* or *refracted* wave), and we see that the change in wavelength causes a sudden change in direction of motion of the wave. If you look carefully you will also see reflected waves which emerge at an angle of reflection equal to the angle of incidence.

Figure (11) shows a beam of yellow light entering a piece of glass. The index of refraction of the glass is 1.55, thus the wavelength of the light in the glass is only .65 times as long as that in air ( $n \approx 1$  for air). You can see both the bending of the ray as it enters the glass and also the reflected ray. (You also see internal reflection and the ray emerging from the bottom surface.) You cannot see the individual wave crests, but otherwise Figures (10) and (11) show similar phenomena.



#### Figure 11

Refraction at surface of glass. When the light waves enter the glass, they travel more slowly and have a shorter wavelength. Like the water waves, the light waves must travel in a different direction in order for the crests to match up.

## **Derivation of Snell's Law**

To calculate the angle by which a light ray is bent when it enters another medium, consider the diagram in Figure (12). The drawing represents a light wave, traveling in a medium of index  $n_1$ , incident on a boundary at an angle  $\theta_1$ . We have sketched successive incident wave crests separated by the wavelength  $\lambda_1$ . Assuming that the index  $n_2$  in the lower medium is greater than  $n_1$ , the wavelength  $\lambda_2$  will be shorter than  $\lambda_1$  and the beam will emerge at the smaller angle  $\theta_2$ .

To calculate the angle  $\theta_2$  at which the transmitted or refracted wave emerges, consider the detailed section of Figure (12) redrawn in Figure (13a). Notice that we have labeled two apparently different angles by the same label  $\theta_1$ . Why these angles are equal is seen in the construction of Figure (13b) where we see that the angles  $\alpha$  and  $\theta_1$  are equal.

## Exercise 2

Show that the two angles labeled  $\,\theta_2\,$  in Figure (13a) must also be equal.



**Figure 12** Analysis of refraction. The crests must match at the boundary between the different wavelength waves.

Since the triangles ACB and ADB are right triangles in Figure (13a), we have

$$\lambda_1 = AB \sin(\theta_1) = \lambda_0/n_1$$
 (3)

$$\lambda_2 = AB \sin(\theta_2) = \lambda_0 / n_2 \tag{4}$$

where AB is the hypotenuse of both triangles and  $\lambda_0$  is the wavelength when  $n_0 = 1$ . When we divide Equation 4 by Equation 5, the distances AB and  $\lambda_0$  cancel, and we are left with

$$\frac{\sin\left(\theta_{1}\right)}{\sin\left(\theta_{2}\right)} = \frac{n_{2}}{n_{1}}$$

or

$$\mathbf{n}_1 \sin\left(\mathbf{\theta}_1\right) = \mathbf{n}_2 \sin\left(\mathbf{\theta}_2\right)$$
 Snells law (5)

Equation 5, known as Snell's law, allows us to calculate the change in direction when a beam of light goes from one medium to another.



# **INTERNAL REFLECTION**

Because of the way rays bend at the interface of two media, there is a rather interesting effect when light goes from a material of higher to a material of lower index of refraction, as in the case of light going from water into air. The effect is seen clearly in Figure (14). Here we have a multiple exposure showing a laser beam entering a tank of water, being reflected by a mirror, and coming out at different angles. The outgoing ray is bent farther away from the normal as it emerges from the water. We reach the point where the outgoing ray bends and runs parallel to the surface of the water. This is a critical angle, for if the mirror is turned farther, the ray can no longer get out and is completely reflected inside the surface.



#### Figure 14

Internal reflection. We took three exposures of a laser beam reflecting off an underwater mirror set at different angles. In the first case the laser beam makes it back out of the water and strikes a white cardboard behind the water tank. In the other two cases, there is total internal reflection at the under side of the water surface. In the final exposure we used a flash to make the mirror visible.



#### Figure 14a

When you are swimming under water and look up, you see the outside world through a round hole. Outside that hole, the surface is a silver mirror. It is easy to calculate the critical angle  $\theta_c$  at which this complete internal reflection begins. Set the angle of refraction,  $\theta_2$  in Figure (14), equal to 90° and we get from Snell's law

$$n_1 \sin \theta_c = n_2 \sin \theta_2 = n_2 \sin 90^\circ = n_2$$
  

$$\sin \theta_c = \frac{n_2}{n_1}; \quad \theta_c = \sin^{-1} \frac{n_2}{n_1} \quad (6)$$

For light emerging from water, we have  $n_2 \approx 1$  for air and  $n_1 = 1.33$  for water giving

$$\sin^{-1}\theta_2 = \frac{1}{1.33} = .75$$
  
$$\theta_c = 48.6^{\circ}$$
(7)

Anyone who swims underwater, scuba divers especially, are quite familiar with the phenomenon of internal reflection. When you look up at the surface of the water, you can see the entire outside world through a circular region directly overhead, as shown in Figure (14a). Beyond this circle the surface looks like a silver mirror.

## Exercise 3

A glass prism can be used as shown in Figure (15) to reflect light at right angles. The index of refraction  $n_g$  of the glass must be high enough so that there is total internal reflection at the back surface. What is the least value  $n_g$  one can have to make such a prism work? (Assume the prism is in the air where  $n \approx 1$ .)



# **Fiber Optics**

Internal reflection plays a critical role in modern communications and modern medicine through fiber optics. When light is sent down through a glass rod or fiber so that it strikes the surface at an angle greater than the critical angle, as shown in Figure (16a), the light will be completely reflected and continue to bounce down the rod with no loss out through the surface. By using modern very clear glass, a fiber can carry a light signal for miles without serious attenuation.

The reason it is more effective to use light in glass fibers than electrons in copper wire for transmitting signals, is that the glass fiber can carry information at a much higher rate than a copper wire, as indicated in Figure (16b). This is because laser pulses traveling through glass, can be turned on and off much more rapidly than electrical pulses in a wire. The practical limit for copper wire is on the order of a million pulses or bits of information per second (corresponding to a *baud rate* of one *megabit*). Typically the information rate is

#### Figure 16a

Because of internal reflections, light can travel down a glass fiber, even when the fiber is bent.



**Figure 16b** A single glass fiber can carry the same amount of information as a fat cable of copper wires.

much slower over commercial telephone lines, not much in excess of 30 to 50 thousand bits of information per second (corresponding to 30 to 50 *kilobaud*). These rates are fast enough to carry telephone conversations or transmit text to a printer, but painfully slow for sending pictures and much too slow for digital television signals. High definition digital television will require that information be sent at a rate of about 3 million bits or pulses every 1/30 of a second for a baud rate of 90 million baud. (Compare that with the baud rate on your computer modem.) In contrast, fiber optics cables are capable of carrying pulses or bits at a rate of about a billion  $(10^9)$  per second, and are thus well suited for transmitting pictures or many phone conversations at once.

By bundling many fine fibers together, as indicated in Figure (17), one can transmit a complete image along the bundle. One end of the bundle is placed up against the object to be observed, and if the fibers are not mixed up, the image appears at the other end.

To transmit a high resolution image, one needs a bundle of about a million fibers. The tiny fibers needed for this are constructed by making a rather large bundle of small glass strands, heating the bundle to soften the glass, and then stretching the bundle until the individual strands are very fine. (If you have heated a glass rod over a Bunsen burner and pulled out the ends, you have seen how fine a glass fiber can be made this way.)



A bundle of glass fibers can be used to carry an image from one point to another. The order of a million fibers are needed to carry the medical images seen on the next page.

# **Medical Imaging**

The use of fiber optics has revolutionized many aspects of medicine. It is an amazing experience to go down and look inside your own stomach and beyond, as the author did a few years ago. This is done with a flexible fiber optics instrument called a retroflexion, producing the results shown in Figure (18). An operation, such as the removal of a gallbladder, which used to require opening the abdomen and a long recovery period, can now be performed through a small hole near the navel, using fiber optics to view the procedure. You can see the viewing instrument and such an operation in progress in Figure (19).





#### Figure 19

Gallbladder operation in progress, being viewed by the rigid laparoscope shown on the right. Such views are now recorded by high resolution television.

# PRISMS

So far in our discussion of refraction, we have considered only beams of light of one color, one wavelength. Because the index of refraction generally changes with wavelength, rays of different wavelength will be bent at different angles when passing the interface of two media. Usually the index of refraction of visible light increases as the wavelength becomes shorter. Thus when white light, which is a mixture of all the visible colors, is sent through a prism as shown in Figure (20), the short wavelength blue light will be deflected by a greater angle than the red light, and the beam of light is separated into a rainbow of colors.



### Figure 20

When light is sent through a prism, it is separated into a rainbow of colors. In this scale drawing, we find that almost all the separation of colors occurs at the second surface where the light emerges from the glass.

# Rainbows

Rainbows in the sky are formed by the reflection and refraction of sunlight by raindrops. It is not, however, particularly easy to see why a rainbow is formed. René Descartes figured this out by tracing rays that enter and leave a spherical raindrop.

In Figure (21a) we have used Snell's law to trace the path of a ray of yellow light that enters a spherical drop of water (of index n = 1.33), is reflected on the back side, and emerges again on the front side. (Only a fraction of the light is reflected at the back, thus the reflected beam is rather weak.) In this drawing, the angle  $\theta_2$  is determined by  $\sin(\theta_1) = 1.33 \sin(\theta_2)$ . At the back, the angles of incidence and reflection are equal, and at the front we have  $1.33 \sin(\theta_2) = \sin(\theta_1)$  (taking the index of refraction of air = 1). Nothing is hard about this construction, it is fairly easy to do with a good drafting program like Adobe Illustrator and a hand calculator.

In Figure (21b) we see what happens when a number of parallel rays enter a spherical drop of water. (This is similar to the construction that was done by Descartes in 1633.) When you look at the outgoing rays, it is not immediately obvious that there is any special direction for the reflected rays. But if you look closely you will see that the ray we have labeled #11 is the one that comes back at the widest angle from the incident ray.

Ray #1, through the center, comes straight back out. Ray #2 comes out at a small angle. The angles increase up to Ray #11, and then start to decrease again for Rays #12 and #13. In our construction the maximum angle, that of Ray #11, was  $41.6^{\circ}$ , close to the theoretical value of  $42^{\circ}$  for yellow light. What is more important than the fact that the maximum angle of deviation is  $42^{\circ}$  is the fact that the rays close to #11 emerge as more or less parallel to each other. The other rays, like those near #3 for example come out at diverging angles. That light is spread out. But the light emerging at  $42^{\circ}$  comes out as a parallel beam. When you have sunlight striking many raindrops, *more yellow light is reflected back at this angle of 42 °than any other angle.* 



**Figure 21c** You will see the yellow part of the rainbow at an angle of 42° as shown above. Red will be seen at a greater angle, blue at a lesser one.

Figure 21a Light ray reflecting from a raindrop.



Repeat the construction for red light where the index of refraction is slightly less than 1.33, and you find the maximum angle of deviation and the direction of the parallel beam is slightly greater than 42°. For blue light, with a higher index, the deviation is less.

If you look at falling raindrops with the sun at your back as shown in Figure (21c), you will see the yellow part of the rainbow along the arc that has an angle of  $42^{\circ}$ from the rays of sun passing you. The red light, having a greater angle of deviation will be above the yellow, and the blue will be below, as you can see in Figure (21d).

Sometimes you will see two or more rainbows if the rain is particularly heavy (we have seen up to 7). These are caused by multiple internal reflections. In the second rainbow there are two internal reflections and the parallel beam of yellow light comes out at an angle of  $51^{\circ}$ . Because of the extra reflection the red is on the inside of the arc and the blue on the outside.

## Exercise 4

Next time you see a rainbow, try to measure the angle the yellow part of the arc makes with the rays of sun passing your head.



Figure 21d Rainbow over Cook's Bay, Moorea.

## **The Green Flash**

The so called green flash at sunset is a phenomenon that is supposed to be very rare, but which is easy to see if you can look at a distant sunset through binoculars. (Don't look until the very last couple of seconds so that you will not hurt your eyes.)

The earth's atmosphere acts as a prism, refracting the light as shown in Figure (22). The main effect is that when you look at a sunset, the sun has already set; only its image is above the horizon. But, as seen in Figure (20), the atmospheric prism also refracts the different colors in the white sunlight at different angles. Due to the fact that the blue light is refracted at a greater angle than the red light, the blue image of the sun is slightly higher above the horizon than the green image, and the green image is higher than the red image. We have over emphasized the displacement of the image in Figure (22). The blue image is only a few percent of the sun's diameter above the red image. Before the sun sets, the various colored images are more or less on top of each other and the sun looks more or less white.

If it is a very clear day, and you watch the sunset with binoculars, just as the sun disappears, for about 1/2 second, the sun turns a deep blue. The reason is that all the other images have set, and for this short time only this blue image is visible. We should call this the "blue flash".



of the white sun as consisting of various colored disks that add up to white. The earth's atmosphere acts as a prism, diffracting the light from the setting sun, separating the colored disks. The blue disk is the last to set. Haze in the atmosphere can block the blue light, leaving the green disk as the last one seen. If the atmosphere is not so clear, if there is a bit of haze or moisture as one often gets in the summer, the blue light is absorbed by the haze, and the last image we see setting is the green image. This is the origin of the green flash. With still more haze you get a red sunset, all the other colors having been absorbed by the haze.

Usually it requires binoculars to see the green or blue colors at the instant of sunset. But sometimes the atmospheric conditions are right so that this final light of the sun is reflected on clouds and can be seen without binoculars. If the clouds are there, there is probably enough moisture to absorb the blue image, and the resulting flash on the clouds is green.

# **Halos and Sun Dogs**

Another phenomenon often seen is the reflection of light from hexagonal ice crystals in the atmosphere. The reflection is seen at an angle of  $22^{\circ}$  from the sun. If the ice crystals are randomly oriented then we get a complete halo as seen in Figure (23a). If the crystals are falling with their flat planes predominately horizontal, we only see the two pieces of the halo at each side of the sun, seen in Figure (24). These little pieces of rainbow are known as "sun dogs".

#### Figure 23

Halo caused by reflection by randomly oriented hexagonal ice crystals.



Figure 24 Sun dogs caused by ice crystals falling flat.



# LENSES

The main impact geometrical optics has had on mankind is through the use of lenses in microscopes, telescopes, eyeglasses, and of course, the human eye. The basic idea behind the construction of a lens is Snell's law, but as our analysis of light reflected from a spherical raindrop indicated, we can get complex results from even simple geometries like a sphere.

Modern optical systems like the zoom lens shown in Figure (25) are designed by computer. Lens design is an ideal problem for the computer, for tracing light rays through a lens system requires many repeated applications of Snell's law. When we analyzed the spherical raindrop, we followed the paths of 12 rays for an index of refraction for only yellow light. A much better analysis would have resulted from tracing at least 100 rays for the yellow index of refraction, and then repeating the whole process for different indices of refraction, corresponding to different wavelengths or colors of light. This kind of analysis, while extremely tedious to do by hand, can be done in seconds on a modern desktop computer.

In this chapter we will restrict our discussion to the simplest of lens systems in order to see how basic instruments, like the microscope, telescope and eye, function. You will not learn here how to design a color corrected zoom lens like the Nikon lens shown below.



Figure 25 Nikon zoom lens.

## **Spherical Lens Surface**

A very accurate spherical surface on a piece of glass is surprisingly easy to make. Take two pieces of glass, put a mixture of grinding powder and water between them, rub them together in a somewhat regular, somewhat irregular, pattern that one can learn in less than 5 minutes. The result is a spherical surface on the two pieces of glass, one being concave and the other being convex. The reason you get a spherical surface from this somewhat random rubbing is that only spherical surfaces fit together perfectly for all angles and rotations. Once the spheres have the desired radius of curvature, you use finer and finer grits to smooth out the scratches, and then jeweler's rouge to polish the surfaces. With any skill at all, one ends up with a polished surface that is perfectly spherical to within a fraction of a wavelength of light.

To see the optical properties of a spherical surface, we can start with the ray diagram we used for the spherical raindrop, and remove the reflections by extending the refracting medium back as shown in Figure (26a). The result is not encouraging. The parallel rays entering near the center of the surface come together—*focus*—quite a bit farther back than rays entering near the outer edge. This range of focal distances is not useful in optical instruments.



**Figure 26a** Focusing properties of a spherical surface. (Not good!)

In Figure (26b) we have restricted the area where the rays are allowed to enter to a small region around the center of the surface. To a very good approximation all these parallel rays come together, focus, at one point. This is the characteristic we want in a simple lens, to bring parallel incoming rays together at one point as the parabolic reflector did.

Figure (26b) shows us that the way to make a good lens using spherical surfaces is to use only the central part of the surface. Rays entering near the axis as in Figure (26b) are deflected only by small angles, angles where we can approximate  $\sin(\theta)$  by  $\theta$  itself. When the angles of deflection are small enough to use small angle approximations, a spherical surface provides sharp focusing. As a result, in analyzing the small angle spherical lenses, we can replace the exact form of Snell's law

$$n_1 \sin(\theta_1) = n_2 \sin(\theta_2)$$
 (5 repeated)

by the approximate equation

$$\mathbf{n}_1 \,\boldsymbol{\theta}_1 = \mathbf{n}_2 \,\boldsymbol{\theta}_2 \qquad \begin{array}{c} Snell's \ law\\ for \ small\\ angles \end{array} \tag{8}$$



#### Figure 26b

We get a much better focus if we use only a small part of the spherical surface.

## Focal Length of a Spherical Surface

Let us now use the simplified form of Snell's law to calculate the focal length f of a spherical surface, i.e., the distance behind the surface where entering parallel rays come to a point. Unless you plan to start making your own lenses, you do not really need this result, but the exercise provides an introduction to how focal lengths are related to the curvature of lenses.

Consider two parallel rays entering a spherical surface as shown in Figure (27). One enters along the axis of the surface, the other a distance h above it. The angle labeled  $\theta_1$  is the angle of incidence for the upper ray, while  $\theta_2$  is the refracted angle. These angles are related by Snell's law

$$\mathbf{n}_1 \, \mathbf{\theta}_1 \, = \, \mathbf{n}_2 \, \mathbf{\theta}_2$$

or

$$\theta_2 = \frac{n_1}{n_2} \theta_1 \tag{9}$$

If you recall your high school trigonometry you will remember that the outside angle of a triangle,  $\theta_1$  in Figure (27a), is equal to the sum of the opposite angles,  $\theta_2$  and  $\alpha$  in this case. Thus

$$\theta_1 = \theta_2 + \alpha$$

or using Equation 9 for  $\theta_2$ 

$$\theta_1 = \frac{n_1}{n_2} \theta_1 + \alpha \tag{10}$$

Now consider the two triangles reproduced in Figures (27b) and (27c). Using the small angle approximation  $\tan(\theta) \approx \sin(\theta) \approx \theta$ , we have for Figure (27b)

$$\theta_1 \approx \frac{\mathbf{h}}{\mathbf{r}} ; \quad \alpha \approx \frac{\mathbf{h}}{f}$$
(11)

Substituting these values for  $\theta_1$  and  $\alpha$  into Equation 10 gives

$$\frac{\mathbf{h}}{\mathbf{r}} = \frac{\mathbf{n}_1}{\mathbf{n}_2}\frac{\mathbf{h}}{\mathbf{r}} + \frac{\mathbf{h}}{f} \tag{12}$$

The height h cancels, and we are left with

$$\frac{1}{f} = \frac{1}{r} \left( 1 - \frac{n_1}{n_2} \right)$$
(13)

The fact that the height h cancels means that parallel rays entering at any height h (as long as the small angle approximation holds) will focus at the same point a distance f behind the surface. This is what we saw in Figure (26b).

Figure (26b) was drawn for  $n_1 = 1$  (air) and  $n_2 = 1.33$ (water) so that  $n_1/n_2 = 1/1.33 = .75$ . Thus for that drawing we should have had

$$\frac{1}{f} = \frac{1}{r} (1 - .75) = \frac{1}{r} (.25) = \frac{1}{r} (\frac{1}{4})$$

$$f = 4r$$
(14)



or

### **Exercise 5**

Compare the prediction of Equation 14 with the results we got in Figure (26b). That is, what do you measure for the relationship between f and r in that figure?

## **Exercise** 6

The index of refraction for red light in water is slightly less than the index of refraction for blue light. Will the focal length of the surface in Figure (26b) be longer or shorter than the focal length for red light?

## Exercise 7

The simplest model for a fixed focus eye is a sphere of index of refraction  $n_2$ . The index  $n_2$  is chosen so that parallel light entering the front surface of the sphere focuses on the back surface as shown in Figure (27d). What value of  $n_2$  is required for this model to work when  $n_1=1$ ? Looking at the table of indexes of refraction, Table 1, explain why such a model would be hard to achieve.

# Aberrations

When parallel rays entering a lens do not come to focus at a point, we say that the lens has an *aberration*. We saw in Figure (26a) that if light enters too large a region of a spherical surface, the focal points are spread out in back. This is called *spherical aberration*. One cure for spherical aberration is to make sure that the diameter of any spherical lens you use is small in comparison to the radius of curvature of the lens surface.



We get rainbows from raindrops and prisms because the index of refraction for most transparent substances changes with wavelength. As we saw in Exercise 6, this causes red light to focus at a different point than yellow or blue light, (resulting in colored bands around the edges of images). This problem is called *chromatic aberration*. The cure for chromatic aberration is to construct complex lenses out of materials of different indices of refraction. With careful design, you can bring the focal points of the various colors back together. Some of the complexity in the design of the zoom lens in Figure (25) is to correct for chromatic aberration.

Astigmatism is a common problem for the lens of the human eye. You get astigmatism when the lens is not perfectly spherical, but is a bit cylindrical. If, for example, the cylindrical axis is horizontal, then light from a horizontal line will focus farther back than light from a vertical line. Either the vertical lines in the image are in focus, or the horizontal lines, but not both at the same time. (In the eye, the cylindrical axis does not have to be horizontal or vertical, but can be at any angle.)

There can be many other aberrations depending upon what distortions are present in the lens surface. We once built a small telescope using a shaving mirror instead of a carefully ground parabolic mirror. The image of a single star stretched out in a line that covered an angle of about 30 degrees. This was an extreme example of an aberration called *coma*. That telescope provided a good example of why optical lenses and mirrors need to be ground very accurately.

What, surprisingly, does not usually cause a serious problem is a small scratch on a lens. You do not get an image of the scratch because the scratch is completely out of focus. Instead the main effect of a scratch is to scatter light and fog the image a bit.

**Figure 27d** *A simple, but hard to achieve, model for an eye.* 

#### Optics-22

Perhaps the most famous aberration in history is the spherical aberration in the primary mirror of the orbiting Hubble telescope. The aberration was caused by an undetected error in the complex apparatus used to test the surface of the mirror while the mirror was being ground and polished. The ironic part of the story is that the aberration could have easily been detected using the same simple apparatus all amateur telescope makers use to test their mirrors (the so called Foucault test), but such a simple minded test was not deemed necessary.

What saved the Hubble telescope is that the engineers found the problem with the testing apparatus, and could therefore precisely determine the error in the shape of the lens. A small mirror, only a few centimeters in diameter, was designed to correct for the aberration in the Hubble image. When this correcting mirror was inserted near the focus of the main mirror, the aberration was eliminated and we started getting the many fantastic pictures from that telescope.

Another case of historical importance is the fact that Issac Newton invented the reflecting telescope to avoid the chromatic aberration present in all lenses at that time. With a parabolic reflecting mirror, all parallel rays entering the mirror focus at a point. The location of the focal point does not depend on the wavelength of the light (as long as the mirror surface is reflecting at that wavelength). You also do not get spherical aberration either because a parabolic surface is the correct shape for focusing, no matter how big the diameter of the mirror is compared to the radius of curvature of the surface.







#### Figure 28

Correction of the Hubble telescope mirror. Top: before the correction. Bottom: same galaxy after correction. Left: astronauts installing correction mirror.
## THIN LENSES

In Figure (29), we look at what happens when parallel rays pass through the two spherical surfaces of a lens. The top diagram (a) is a reproduction of Figure (26b) where a narrow bundle of parallel rays enters a new medium through a single spherical surface. By making the diameter of the bundle of rays much less than the radius of curvature of the surface, the parallel rays all focus to a single point. We were able to calculate where this point was located using small angle approximations.

In Figure (29b), we added a second spherical surface. The diagram is drawn to scale for indices of refraction n = 1 outside the gray region and n = 1.33 inside, and using Snell's law at each interface of each ray. (The drawing program Adobe Illustrator allows you to do this quite accurately.) The important point to note is that the parallel rays still focus to a point. The difference is that the focal point has moved inward.



Figure 29

A two surface lens. Adding a second surface still leaves the light focused to a point, as long as the diameter of the light bundle is small compared to the radii of the lens surfaces. In Figure (29c), we have moved the two spherical surfaces close together to form what is called a *thin lens*. We have essentially eliminated the distance the light travels between surfaces. If the index of refraction outside the lens is 1 and has a value n inside, and surfaces have radii of curvature  $r_1$  and  $r_2$ , then the focal length *f* of the lens given by the equation

$$\boxed{\frac{1}{f} = \left(n-1\right)\left(\frac{1}{r_1} + \frac{1}{r_2}\right)} \qquad \substack{lens \ maker's \\ equation} \qquad (15)$$

Equation 15, which is known as the *lens maker's equation*, can be derived in a somewhat lengthy exercise involving similar triangles.

Unless you are planning to grind your own lenses, the lens maker's equation is not something you will need to use. When you buy a lens, you specify what focal length you want, what diameter the lens should be, and whether or not it needs to be corrected for color aberration. You are generally not concerned with how the particular focal length was achieved—what combination of radii of curvatures and index of refraction were used.

## **Exercise 8**

(a) See how well the lens maker's equation applies to our scale drawing of Figure (29c). Our drawing was done to a scale where the spherical surfaces each had a radius of  $r_1 = r_2 = 37$  mm, and the distance *f* from the center of the lens to the focal point was 55 mm.

(b) What would be the focal length *f* of the lens if it had been made from diamond with an index of refraction n = 2.42?

# **The Lens Equation**

What is important in the design of a simple lens system is where images are formed for objects that are different distances from the lens. Light from a very distant object enters a lens as parallel rays and focuses at a distance equal to the focal length f behind the lens. To locate the image when the object is not so far away, you can either use a simple graphical method which involves a tracing of two or three rays, or use what is called the *lens equation* which we will derive shortly from the graphical approach.

For our graphical work, we will use an arrow for the object, and trace out rays coming from the tip of the arrow. Where the rays come back together is where the image is formed. We will use the notation that the object is at a distance (o) from the lens, and that the image is at a distance (i) as shown in Figure (30).

In Figure (30) we have located the image by tracing three rays from the tip of the object. The top ray is parallel to the axis of the lens, and therefore must cross the axis at the focal point behind the lens. The middle



## Figure 30

Locating the image using ray tracing. Three rays are easy to draw. One ray goes straight through the center of the lens. The top ray, parallel to the axis, intersects the axis where parallel rays would focus. A ray going through the left focus, comes out parallel to the axis. The image of the arrow tip is located where these rays intersect. ray, which goes through the center of the lens, is undeflected if the lens is thin. The bottom ray goes through the focal point in front of the lens, and therefore must come out parallel to the axis behind the lens. (Lenses are symmetric in that parallel light from either side focuses at the same distance f from the lens.) The image is formed where the three rays from the tip merge. To locate the image, you only need to draw two of these three special rays.

## **Exercise 9**

(a) Graphically locate the image of the object in Figure (31).

(b) A ray starts out from the tip of the object in the direction of the dotted line shown. Trace out this ray through the lens and show where it goes on the back side of the lens.

In Exercise 9, you found that, once you have located the image, you can trace out any other ray from the tip of the object that passes through the lens, because these rays must all pass through the tip of the image.



## Figure 31

Locate the image of the arrow, and then trace the ray starting out in the direction of the dotted line.

There is a very, very simple relationship between the object distance *o*, the image distance *i* and the lens focal length *f*. It is

$$\frac{1}{o} + \frac{1}{i} = \frac{1}{f} \qquad the lens \\ equation \qquad (16)$$

Equation 16 is worth memorizing if you are going to do any work with lenses. It is the equation you will use all the time, it is easy to remember, and as you will see now, the derivation requires some trigonometry you are not likely to remember. We will take you through the derivation anyway, because of the importance of the result.

In Figure (32a), we have an object of height A that forms an inverted image of height B. We located the image by tracing the top ray parallel to the axis that passes through the focal point behind the lens, and by tracing the ray that goes through the center of the lens.





In Figure (32b) we have selected one of the triangles that appears in Figure (32a). The triangle starts at the tip of the object, goes parallel to the axis over to the image, and then down to the tip of the image. The length of the triangle is (o + i) and the height of the base is (A+B). The lens cuts this triangle to form a smaller similar triangle whose length is o and base is (A). The ratio of the base to length of these similar triangles must be equal, giving

$$\frac{\mathbf{A}}{o} = \frac{\mathbf{A} + \mathbf{B}}{\left(o + i\right)} \Rightarrow \frac{\left(\mathbf{A} + \mathbf{B}\right)}{\mathbf{A}} = \frac{\left(o + i\right)}{o} \tag{17}$$

In Figure (32c) we have selected another triangle which starts where the top ray hits the lens, goes parallel to the axis over to the image, and down to the tip of the image. This triangle has a length i and a base of height (A+B) as shown. This triangle is cut by a vertical line at the focal plane, giving a smaller similar triangle of length f and base (A) as shown. The ratio of the length to base of these similar triangles must be equal, giving

$$\frac{A}{f} = \frac{A+B}{i} \Rightarrow \frac{(A+B)}{A} = \frac{i}{f}$$
(18)

Combining Equations 17 and 18 gives

$$\frac{i}{f} = \frac{o+i}{o} = 1 + \frac{i}{o} \tag{19}$$

Finally, divide both sides by *i* and we get

$$\frac{1}{f} = \frac{1}{i} + \frac{1}{o} \qquad \begin{array}{c} lens\\ equation \end{array}$$
(16)

which is the lens equation, as advertised.

Note that the lens equation is an exact consequence of the geometrical construction shown back in Figure (30). There is no restriction about small angles. However if you are using spherical lenses, you have to stick to small angles or the light will not focus to a point.

## **Negative Image Distance**

The lens equation is more general than you might expect, for it works equally well for positive and negative distances and focal lengths. Let us start by seeing what we mean by a negative image distance. Writing Equation 15 in the form

$$\frac{1}{i} = \frac{1}{f} - \frac{1}{o} \tag{16a}$$

let us see what happens if 1/o is bigger than 1/f so that *i* turns out to be negative. If 1/o is bigger than 1/f, that means that *o* is less than *f* and we have placed the object within the focal length as shown in Figure (33).

When we trace out two rays from the tip of the image, we find that the rays diverge after they pass through the lens. They diverge as if they were coming from a point behind the object, a point shown by the dotted lines. In this case we have what is called a *virtual image*, which is located at a *negative image distance* (*i*). This negative image distance is correctly given by the lens equation (16a).

(We will not drag you through another geometrical proof of the lens equation for negative image distances. It should be fairly convincing that just when the image distance becomes negative in the lens equation, the geometry shows that we switch from a real image on the right side of the lens to a virtual image on the left.)



**Figure 33** When the object is located within the focal length, we get a virtual image behind the object.

# Negative Focal Length and Diverging Lenses

In Figure (33) we got a virtual image by moving the object inside the focal length. Another way to get a virtual image is to use a diverging lens as shown in Figure (34). Here we have drawn the three special rays, but the role of the focal point is reversed. The ray through the center of the lens goes through the center as before. The top ray parallel to the axis of the lens diverges outward as if it came from the focal point on the left side of the lens. The ray from the tip of the object headed for the right focal point, comes out parallel to the axis. Extending the diverging rays on the right, back to the left side, we find a virtual image on the left side.

You get diverging lenses by using concave surfaces as shown in Figure (34). In the lens maker's equation,

$$\frac{1}{f} = \left(n - 1\right) \left(\frac{1}{r_1} + \frac{1}{r_2}\right) \qquad \begin{array}{c} lensmaker's \\ equation \end{array}$$
(15)

you replace 1/r by - 1/r for any concave surface. If 1/f turns out negative, then you have a diverging lens. Using this negative value of f in the lens equation (with f = -|f|) we get

$$\frac{1}{i} = -\left(\frac{1}{\left|f\right|} + \frac{1}{o}\right) \tag{16b}$$

This always gives a negative image distance *i*, which means that diverging lenses only give virtual images.



A diverging lens always gives a virtual image.

## **Exercise 10**

You have a lens making machine that can grind surfaces, either convex or concave, with radius of curvatures of either 20 cm or 40 cm, or a flat surface. How many different kinds of lenses can you make? What is the focal length and the name of the lens type for each lens? Figure (35) shows the names given to the various lens types.

## **Negative Object Distance**

With the lens equation, we can have negative image distances and negative focal lengths, and also negative object distances as well.

In all our drawings so far, we have drawn rays coming out of the tip of an object located at a positive object distance. A negative object distance means we have a virtual object where rays are converging toward the tip of the virtual object but don't get there. A comparison of the rays emerging from a real object and converging toward a virtual object is shown in Figure (36). The converging rays (which were usually created by some other lens) can be handled with the lens equation by assuming that the distance from the lens to the virtual object is negative.





Various lens types. Note that eyeglasses are usually meniscus convex or meniscus concave.

As an example, suppose we have rays converging to a point, and we insert a diverging lens whose negative focal length f = -|f| is equal to the negative object distance o = -|o| as shown in Figure (37). The lens equation gives

$$\frac{1}{i} = \frac{1}{f} - \frac{1}{o} = \frac{1}{-|f|} - \frac{1}{-|o|} = \frac{1}{|o|} - \frac{1}{|f|}$$
(20)

If |f| = |o|, then 1/i = 0 and the image is infinitely far away. This means that the light emerges as a parallel beam as we showed in Figure (37).



**Figure 36** *Positive and negative object distances.* 



**Figure 37** *Negative focal length.* 

## **Multiple Lens Systems**

Using the lens equation, and knowing how to handle both positive and negative distances and focal lengths, you can design almost any simple lens system you want. The idea is to work your way through the system, one lens at a time, where the image from one lens becomes the object for the next. We will illustrate this process with a few examples.

As our first example, consider Figure (38a) where we have two lenses of focal lengths  $f_I = 10$  cm and  $f_2 = 12$  cm separated by a distance D = 40 cm. An object placed at a distance  $o_I = 17.5$  cm from the first lens creates an image a distance  $i_1$  behind the first lens. Using the lens equation, we get

$$i_I = \frac{1}{f_I} - \frac{1}{o_I} = \frac{1}{10} - \frac{1}{17.5} = \frac{1}{23.33}$$
 (21)

 $i_1 = 23.33$  cm

the same distance we got graphically in Figure (38a).

This image, which acts as the object for the second lens has an object distance

$$o_2 = D - i_1 = 40 \text{ cm} - 23.33 \text{ cm} = 16.67 \text{ cm}$$

This gives us a final upright image at a distance  $i_2$  given by

$$\frac{1}{i_2} = \frac{1}{f_2} - \frac{1}{o_2} = \frac{1}{12} - \frac{1}{16.67} = \frac{1}{42.86}$$
(22)

$$i_2 = 42.86 \text{ cm}$$
 (23)

which also accurately agrees with the geometrical construction.



**Figure 38a** Locating the image in a two lens system.

In Figure (38b), we moved the second lens up to within 8 cm of the first lens, so that the first image now falls behind the second lens. We now have a negative object distance

$$o_2 = D - i_1 = 8 \text{ cm} - 23.33 \text{ cm} = -15.33 \text{ cm}$$

Using this negative object distance in the lens equation gives

$$\frac{1}{i_2} = \frac{1}{f_2} - \frac{1}{o_2} = \frac{1}{12} - \frac{1}{-15.33}$$
$$= \frac{1}{12} + \frac{1}{15.33} = \frac{1}{6.73}$$
$$i_2 = 6.73 \text{ cm}$$
(24)

In the geometrical construction we find that the still inverted image is in fact located 6.73 cm behind the second image.

While it is much faster to use the lens equation than trace rays, it is instructive to apply both approaches for a few examples to see that they both give the same result. In drawing Figure (38b) an important ray was the one that went from the tip of the original object, down through the first focal point. This ray emerges from the first lens traveling parallel to the optical axis. The ray then enters the second lens, and since it was parallel to the axis, it goes up through the focal point of the second lens as shown. The second image is located by drawing the ray that passes straight through the second lens, heading for the tip of the first image. Where these two rays cross is where the tip of the final image is located.



#### Figure 38b

We moved the second lens in so that the second object distance is negative. We now get an inverted image 6.73 cm from the second lens. In Figure (38c) we sketched a number of rays passing through the first lens, heading for the first image. These rays are converging on the second lens, which we point out in Figure (36b) was the condition for a negative object distance.



Figure 38c

## **Two Lenses Together**

If you put two thin lenses together, as shown in Figure (39), you effectively create a new thin lens with a different focal length. To find out what the focal length of the combination is, you use the lens equation twice, setting the second object distance  $o_2$  equal to minus the first image distance  $-i_1$ .

$$o_2 = -i_1 \qquad for two lenses \\ together$$
 (25)

From the lens equations we have

$$\frac{1}{i_I} = \frac{1}{f_I} - \frac{1}{o_I}$$
(26)

$$\frac{1}{i_2} = \frac{1}{f_2} - \frac{1}{o_2} \tag{27}$$

Setting  $o_2 = -i_1$  in Equation 27 gives



Two lenses together. Since the object for the second lens is on the wrong side of the lens, the object distance  $o_2$  is negative in this diagram. If the lenses are close together,  $i_1$  and  $-o_2$  are essentially the same.

Using Equation 26 for  $1/i_1$  gives

$$\frac{1}{i_2} = \frac{1}{f_2} + \frac{1}{f_1} - \frac{1}{o_1}$$
$$\frac{1}{o_1} + \frac{1}{i_2} = \frac{1}{f_1} + \frac{1}{f_2}$$
(28)

Now  $o_1$  is the object distance and  $i_2$  is the image distance for the pair of lenses. Treating the pair of lenses as a single lens, we should have

$$\frac{1}{o_1} + \frac{1}{i_2} = \frac{1}{f}$$
(29)

where f is the focal length of the combined lens.

Comparing Equations 28 and 29 we get

$$\frac{1}{f} = \frac{1}{f_1} + \frac{1}{f_2} \qquad \begin{array}{c} \text{focal length of two} \\ \text{thin lenses together} \end{array} (30)$$

as the simple formula for the combined focal length.

## Exercise 11

(a) Find the image distances  $i_2$  for the geometry of Figures (38), but with the two lenses reversed, i.e., with  $f_1 = 12 \text{ cm}$ ,  $f_2 = 10 \text{ cm}$ . Do this for both length D = 40 cm and D = 8 cm.

(b) If the two lenses are put together (D = 0) what is the focal length of the combination?

# Magnification

It is natural to define the magnification created by a lens as the ratio of the height of the image to the height of the object. In Figure (40) we have reproduced Figure (38a) emphasizing the heights of the objects and images.

We see that the shaded triangles are similar, thus the ratio of the height B of the first image to the height A of the object is

$$\frac{B}{A} = \frac{i_l}{o_l} \tag{31}$$

We could define the magnification in the first lens as the ratio of B/A, but instead we will be a bit tricky and include a - (minus) sign to represent the fact that the image is inverted. With this convention we get

$$m_{I} = \frac{-\mathbf{B}}{\mathbf{A}} = \frac{-i_{I}}{o_{I}} \qquad \begin{array}{c} definition \ of \\ magnification \ m \end{array} \tag{32}$$

Treating B as the object for the second lens gives

$$m_2 = \frac{-C}{B} = \frac{-i_2}{o_2}$$
(33)

The total magnification  $m_{12}$  in going from the object A to the final image C is

$$m_{12} = \frac{C}{A} \tag{34}$$

which has a + sign because the final image C is upright. But

$$\frac{C}{A} = \left(\frac{-C}{B}\right)\left(\frac{-B}{A}\right)$$
(35)

Thus we find that the final magnification is the product of the magnifications of each lens.



**Figure 40** *Magnification of two lenses.* 

## Exercise 12

Figures (38) and (40) are scale drawings, so that the ratio of image to object sizes measured from these drawings should equal the calculated magnifications.

(a) Calculate the magnifications  $m_1$ ,  $m_2$  and  $m_{12}$  for Figure (38a) or (40) and compare your results with magnifications measured from the figure.

(b) Do the same for Figure (38b). In Figure (38b), the final image is inverted. Did your final magnification  $m_{12}$  come out negative?

## Exercise 13

Figure (41a) shows a magnifying glass held 10 cm above the printed page. Since the object is inside the focal length we get a virtual image as seen in the geometrical construction of Figure (41b). Show that our formulas predict a positive magnification, and estimate the focal length of the lens. (Answer: about 17 cm.)



Figure 41a Using a magnifying glass.



Figure 41b When the magnifying glass is le

When the magnifying glass is less than a focal length away from the object, we see an upright virtual image.

# THE HUMAN EYE

A very good reason for studying geometrical optics is to understand how your own eye works, and how the situation is corrected when something goes wrong.

Back in Exercise 6 (p21), during our early discussion of spherical lens surfaces, we considered as a model of an eye a sphere of index of refraction  $n_2$ , where  $n_2$ was chosen so that parallel



rays which entered the front surface focused on the back surface as shown in Figure (27d). The value of  $n_2$  turned out to be  $n_2 = 2.0$ . Since the only common substance with an index of refraction greater than zircon at n = 1.923 is diamond at n = 2.417, it would be difficult to construct such a model eye. Instead some extra focusing capability is required, both to bring the focus to the back surface of the eye, and to focus on objects located at various distances.



## Figure 42

The human eye. The cornea and the lens together provide the extra focusing power required to focus light on the retina. (Photograph of the human eye by Lennart Nilsson.)



Figure (42a) is a sketch of the human eye and Figure 42b a remarkable photograph of the eye. As seen in (42a), light enters the *cornea* at the front of the eye. The amount of light allowed to enter is controlled by the opening of the *iris*. Together the cornea and *crystalline lens* focuses light on the retina which is a film of *nerve fibers* on the back surface of the eye. Information from the new fibers is carried to the brain through the optic nerve at the back. In the retina there are two kinds of nerve fibers, called *rods* and *cones*. Some of the roughly 120 million rods and 7 million cones are seen magnified about 5000 times in Figure (43). The slender ones, the rods, are more sensitive to dim light, while the shorter, fatter, cones, provide our color sensitivity.

In our discussion of the human ear, we saw how there was a mechanical system involving the basilar membrane that distinguished between the various frequencies of incoming sound waves. Information from nerves attached to the basilar membrane was then enhanced through processing in the local nerve fibers before being sent to the brain via the auditory nerve. In the eye, the nerve fibers behind the retina, some of which can be seen on the right side of Figure (43), also do a considerable amount of information processing before the signal travels to the brain via the optic nerve. The way that information from the rods and cones is processed by the nerve fibers is a field of research.

Returning to the front of the eye we have the surface of the cornea and the crystalline lens focusing light on the retina. Most of the focusing is done by the cornea. The shape, and therefore the focal length of the crystalline lens can be altered slightly by the *ciliary muscle* in order to bring into focus objects located at different distances.

rod Figure 43 Rods and cones in the retina. The thin ones are the rods, the fat ones the cones.



In a normal eye, when the ciliary muscle is in its resting position, light from infinity is focused on the retina as shown in Figure (44a). To see a closer object, the ciliary muscles contract to shorten the focal length of the cornea-lens system in order to continue to focus light on the cornea (44b). If the object is too close as in Figure (44c), the light is no longer focused and the object looks blurry. The shortest distance at which the light remains in focus is called the *near point*. For children the near point is as short as 7 cm, but as one ages and the crystalline lens becomes less flexible, the near point recedes to something like 200 cm. This is why older people hold written material far away unless they have reading glasses.



Parallel light rays from a distant object are focuses on the retina when the ciliary muscles are in the resting position.



## Figure 44b

The ciliary muscle contracts to shorten the focal length of the cornea-lens system in order to focus light from a more nearby object.



## Figure 44c

When an object is to close, the light cannot be focused. The closest distance we. ....

# **Nearsightedness and Farsightedness**

Not all of us have the so called *normal* eyes described by Figure (44). There is increasing evidence that those who do a lot of close work as children end up with a condition called *nearsightedness* or *myopia* where the eye is elongated and light from infinity focuses inside the eye as shown in Figure (45a). This can be corrected by placing a diverging lens in front of the eye to move the focus back to the retina as shown in Figure (45b).

The opposite problem, farsightedness, where light focuses behind the retina as shown in Figure (46a) is corrected by a converging lens as shown in Figure (46b).



**Figure 45** *Nearsightedness can be corrected by a convex lens..* 



Farsightedness can be corrected by a convex lens

# THE CAMERA

There are a number of similarities between the human eye and a simple camera. Both have an iris to control the amount of light entering, and both record an image at the focal plane of the lens. In a camera, the focus is adjusted, not by changing the shape of the lens as in the eye, but by moving the lens back and forth. The eye is somewhat like a TV camera in that both record images at a rate of about 30 per second, and the information is transmitted electronically to either the brain or a TV screen.

On many cameras you will find a series of numbers labeled by the letter f, called the f number or f stop. Just as for the parabolic reflectors in figure 4 (p5), the fnumber is the ratio of the lens focal length to the lens diameter. As you close down the iris of the camera to reduce the amount of light entering, you reduce the effective diameter of the lens and therefore increase the f number.

## **Exercise 14**

The iris on the human eye can change the diameter of the opening to the lens from about 2 to 8 millimeters. The total distance from the cornea to the retina is typically about 2.3 cm. What is the range of f values for the human eye? How does this range compare with the range of f value on your camera? (If you have one of the automatic point and shoot cameras, the f number and the exposure time are controlled electronically and you do not get to see or control these yourself.)



**Figure 47a** The Physics department's Minolta single lens reflex camera.



## Figure 47b

The lens system for a Nikon single lens reflex camera. When you take the picture, the hinged mirror flips out of the way and the light reaches the film. Before that, the light is reflected through the prism to the eyepiece.

# **Depth of Field**

There are three ways to control the exposure of the film in a camera. One is by the speed of the film, the second is the exposure time, and the third is the opening of the iris or f stop. In talking a picture you should first make sure the exposure is short enough so that motion of the camera and the subject do not cause blurring. If your film is fast enough, you can still choose between a shorter exposure time or a smaller f stop. This choice is determined by the *depth of field* that you want.

The concept of depth of field is illustrated in Figures (48a and b). In (48a), we have drawn the rays of light from an object to an image through an f2 lens, a lens with a focal length equal to twice its diameter. (The effective diameter can be controlled by a flexible diaphragm or iris like the one shown.) If you placed a film at the image distance, the point at the tip of the object arrow would focus to a point on the film. If you moved the film forward to position 1, or back to position 2, the image of the arrow tip would fill a circle about equal to the thickness of the three rays we drew in the diagram.

If the film were ideal, you could tell that the image at positions 1 or 2 was out of focus. But no film or recording medium is ideal. If you look closely enough there is always a graininess caused by the size of the basic medium like the silver halide crystals in black and white film, the width of the scan lines in an analog TV camera, or the size of the pixels in a digital camera. If the image of the arrow tip at position 1 is smaller than the grain or pixel size then you cannot tell that the picture is out of focus. You can place the recording medium anywhere between position 1 and 2 and the image will be as sharp as you can get.

In Figure (48b), we have drawn the rays from the same object passing through a smaller diameter f8 lens. Again we show by dotted lines positions 1 and 2 where the image of the arrow point would fill the same size circle as it did at positions 1 and 2 for the f2 lens above. Because the rays from the f8 lens fill a much narrower cone than those from the f2 lens, there is a much greater distance between positions 1 and 2 for the f8 lens.



A large diameter lens has a narrow depth of field.



Photograph taken at f 5.6.



Photograph taken at f 22.

## Figure 48b

Reducing the effective diameter of the lens increases the depth of field.

If Figures (48) represented a camera, you would not be concerned with moving the film back and forth. Instead you would be concerned with how far the image could be moved back and forth and still appear to be in focus. If the film were at the image position and you then moved the object in and out, you could not move it very far before it's image was noticeably out of focus with the f2 lens. You could move it much farther for the f8 lens.

This effect is illustrated by the photographs on the right side of Figures 48, showing a close-up tree and the distant tower on Baker Library at Dartmouth College. The upper picture taken at f 5.6 has a narrow depth of field, and the tower is well out of focus. In the bottom picture, taken at f 22, has a much broader depth of field and the tower is more nearly in focus. (In both cases we focused on the nearby tree bark.)

Camera manufacturers decide how much blurring of the image is noticeable or tolerable, and then figure out the range of distances the object can be moved and still be acceptably in focus. This range of distance is called the *depth of field*. It can be very short when the object is up close and you use a wide opening like f2. It can be quite long for a high f number like f22. The inexpensive fixed focus cameras use a small enough lens so that all objects are "in focus" from about 3 feet or 1 meter to infinity.

In the extreme limit when the lens is very small, the depth of field is so great that everything is in focus



Figure 48c

Camera lens. This lens is set to f11, and adjusted to a focus of 3 meters or 10 ft. At this setting, the depth of field ranges from 2 to 5 meters.

everywhere behind the lens. In this limit you do not even need a lens, a pinhole in a piece of cardboard will do. If enough light is available and the subject doesn't not move, you can get as good a picture with a pinhole camera as one with an expensive lens system. Our pinhole camera image in Figure (49) is a bit fuzzy because we used too big a pinhole.

(If you are nearsighted you can see how a pinhole camera works by making a tiny hole with your fingers and looking at a distant light at night without your glasses. Just looking at the light, it will look blurry. But look at the light through the hole made by your fingers and the light will be sharp. You can also see the eye chart better at the optometrists if you look through a small hole, but they don't let you do that.)



**Figure 49a** We made a pinhole camera by replacing the camera lens with a plastic film case that had a small hole poked into the end.



## Figure 49b

Photograph of Baker library tower, taken with the pinhole camera above. If we had used a smaller hole we would have gotten a sharper focus.

# Eye Glasses and a Home Lab Experiment

When you get a prescription for eyeglasses, the optometrist writes down number like -1.5, -1.8 to represent the *power* of the lenses you need. These cryptic number are the power of the lenses measured in *diopters*. What a diopter is, is simply the reciprocal of the focal length 1/f, where *f* is measured in meters. A lens with a power of 1 diopter is a converging lens with a focal length of 1 meter. Those of us who have lenses closer to -4 in power have lenses with a focal length of -25 cm, the minus sign indicating a diverging lens to correct for nearsightedness as shown back in Figure (45).

If you are nearsighted and want to measure the power of your own eyeglass lenses, you have the problem that it is harder to measure the focal length of a diverging lens than a converging lens. You can quickly measure the focal length of a converging lens like a simple magnifying glass by focusing sunlight on a piece of paper and measuring the distance from the lens to where the paper is starting to smoke. But you do not get a real image for a diverging lens, and cannot use this simple technique for measuring the focal length and power of diverging lenses used by the nearsighted.

As part of a project, some students used the following method to measure the focal length and then determine the power in diopters, of their and their friend's eye-glasses. They started by measuring the focal length  $f_0$  of a simple magnifying glass by focusing the sun. Then

they placed the magnifying glass and the eyeglass lens together, measured the focal length of the combination, and used the formula

$$\frac{1}{f} = \frac{1}{f_1} + \frac{1}{f_2}$$
 (30 repeated)

to calculate the focal length of the lens.

(Note that if you measure distances in meters, then  $1/f_1$  is the power of lens 1 in diopters and  $1/f_2$  that of lens 2. Equation 30 tells you that the power of the combination 1/f is the sum of the powers of the two lenses.

## Exercise 15

Assume that you find a magnifying lens that focuses the sun at a distance of 10 cm from the lens. You then combine that with one of your (or a friends) eyeglass lenses, and discover that the combination focus at a distance of 15 cm. What is the power, in diopters, of

- (a) The magnifying glass.
- (b) The combination.
- (c) The eyeglass lens.

## Exercise 16 – Home Lab

Use the above technique to measure the power of your or your friend's glasses. If you have your prescription compare your results with what is written on the prescription. (The prescription will also contain information about axis and amount of astigmatism. That you cannot check as easily.

# THE EYEPIECE

When the author was a young student, he wondered why you do not put your eye at the focal point of a telescope mirror. That is where the image of a distance object is, and that is where you put the film in order to record the image. You do not put your eye at the image because it would be like viewing an object by putting your eyeball next to it. The object would be hopelessly out of focus. Instead you look through an eyepiece.

The eyepiece is a magnifying glass that allows your eye to comfortably view an image or small object up close. For a normal eye, the least eyestrain occurs when looking at a distant object where the light from the object enters the eye as parallel rays. It is then that the ciliary muscles in the eye are in a resting position. If the image or small object is placed at the focal plane of a lens, as shown in Figure (50), light emerges from the lens as parallel rays. You can put your eye right up to that lens, and view the object or image as comfortably as you would view a distant scene.



## Figure 50

The eyepiece or magnifier. To look at small object, or to study the image produced by another lens or mirror, place the image or object at the focal plane of a lens, so that the light emerges as parallel rays that your eye can comfortably focus upon.

## **Exercise 17 - The Magnifying Glass**

There are three distinct ways of viewing an object through a magnifying glass, which you should try for yourself. Get a magnifying glass and use the letters on this page as the object to be viewed.

(a) First measure the focal length of the lens by focusing the image of a distant object onto a piece of paper. A light bulb across the room or scene out the window will do.

(b) Draw some object on the paper, and place the paper at least several focal lengths from your eye. Then hold the lens about 1/2 a focal length above the object as shown in Figure (51a). You should now see an enlarged image of the object as indicated in Figure (51a). You are now looking at the virtual image of the object. Check that the magnification is roughly a factor of  $2\times$ .

(c) Keeping your eye in the same position, several focal lengths and at least 20 cm from the paper, pull the lens back toward your eye. The image goes out of focus when the lens is one focal length above the paper, and then comes back into focus upside down when the lens is farther out. You are now looking at the real image as indicated in Figure (51b). Keep your head far enough back that your eye can focus on this real image.

Hold the lens two focal lengths above the page and check that the inverted real image of the object looks about the same size as the object itself. (As you can see from Figure (51b), the inverted image should be the same size as the object, but 4 focal lengths closer.)





## Optics-38

(d) Now hold the lens one focal length above the page and put your eye right up to the lens. You are now using the lens as an eyepiece as shown in Figure (50). The letters will be large because your eye is close to them, and they will be comfortably in focus because the rays are entering your eye as parallel rays like the rays from a distant object. When you use the lens as an eyepiece you are not looking at an image as you did in parts (b) and (c) of this exercise, instead your eye is creating an image on your retina from the parallel rays.

(e) As a final exercise, hold the lens one focal length above a page of text, start with your eye next to the lens, and then move your head back. Since the light from the page is emerging from the lens as parallel rays, the size of the letters should not change as you move your head back. Instead what you should see is fewer and fewer letters in the magnifying glass as the magnifying glass itself looks smaller when farther away. This effect is seen in Figure (52).



## Figure 52

When the lens is one focal length from the page, the emerging rays are parallel. Thus the image letters do not change size as we move away. Instead the lens looks smaller, and we see fewer letters in the lens.

## **The Magnifier**

When jewelers work on small objects like the innards of a watch, they use what they call a *magnifier* which can be a lens mounted at one end of a tube as shown in Figure (53). The length of the tube is equal to the focal length of the lens, so that if you put the other end of the tube up against an object, the lens acts as an eyepiece and light from the object emerges from the lens as parallel rays. By placing your eye close to the lens, you get a close up, comfortably seen view of the object. You may have seen jewelers wear magnifiers like that shown in Figure (54).



**Figure 53** *A magnifier*.



**Figure 54** Jeweler Paul Gross with magnifier lenses mounted in visor.

## **Angular Magnification**

Basically all the magnifier does is to allow you to move the object close to your eye while keeping the object comfortably in focus. It is traditional to define the *magnification* of the magnifier as the ratio of the size of the object as seen through the lens to the size of the object as you would see it without a magnifier. By size, we mean the angle the object subtends at your eye. This is often called the *angular magnification*.

The problem with this definition of magnification is that different people, would hold the object at different distances in order to look at it without a magnifier. For example, us nearsighted people would hold it a lot closer than a person with normal vision. To avoid this ambiguity, we can choose some standard distance like 25 cm, a standard near point, at which a person would normally hold an object when looking at it. Then the angular magnification of the magnifier is the ratio of the angle  $\theta_m$  subtended by the object when using the magnifier, as shown in Figure (55a), to the angle  $\theta_0$ subtended by the object held at a distance of 25 cm, as shown in Figure (55b).

angular magnification =  $\frac{\theta_{\rm m}}{\theta_0}$  angles defined in Figure 55 (37) To calculate the angular magnification we use the small angle approximation  $\sin \theta \approx \theta$  to get

$$\Theta_{\rm m} = \frac{y}{f}$$
 from Figure 55a  
 $\Theta_0 = \frac{y}{25 \, \rm cm}$  from Figure 55b

which gives

$$\frac{\text{angular}}{\text{magnification}} = \frac{y/f}{y/25 \text{ cm}} = \frac{25 \text{ cm}}{f}$$
(38)

Thus if our magnifier lens has a focal length of 5 cm, the angular magnification is  $5\times$ . Supposedly the object will look five times bigger using the magnifier than without it.



**Figure 55** *The angles used in defining angular magnification.* 

# TELESCOPES

The basic design of a telescope is to have a large lens or parabolic mirror to create a bright real image, and then use an eyepiece to view the image. If we use a large lens, that lens is called an *objective lens*, and the telescope is called a *refracting telescope*. If we use a parabolic mirror, then we have a *reflecting telescope*.

The basic design of a refracting telescope is shown in Figure (56). Suppose, as shown in Figure (56a), we are looking at a constellation of stars that subtend an angle  $\theta_0$  as viewed by the unaided eye. The eye is directed just below the bottom star and light from the top star enters at an angle  $\theta_0$ . In Figure (56b), the lens system from the telescope is placed in front of the eye, and we are following the path of the light from the top star in the constellation.

The parallel rays from the top star are focused at the focal length  $f_0$  of the objective lens. We adjust the eyepiece so that the image produced by the objective lens is at the focal point of the eyepiece lens, so that light from the image will emerge from the eyepiece as parallel rays that the eye can easily focus.

## Figure 56a

The unaided eye looking at a constellation of stars that subtend an angle  $\theta_o$ .

As with the magnifier, we define the magnification of the telescope as the ratio of the size of (angle subtended by) the object as seen through the object to the size of (angle subtended by) the object seen by the unaided eye. In Figure (56) we see that the constellation subtends an angle  $\theta_0$  as viewed by the unaided eye, and an angle  $\theta_i$  when seen through the telescope. Thus we define the magnification of the telescope as

$$m = \frac{\theta_i}{\theta_0} \qquad \begin{array}{c} magnification \\ of telescope \end{array} \tag{39}$$

To calculate this ratio, we note from Figure (56c) that, using the small angle approximation  $\sin\theta \approx \theta$ , we have

$$\theta_0 = \frac{y_i}{f_0} ; \quad \theta_i = \frac{y_i}{f_e}$$
(40)

where  $f_0$  and  $f_e$  are the focal lengths of the objective and eyepiece lens respectively. In the ratio, the image height  $y_i$  cancels and we get



## Figure 56b

Looking at the same constellation through a simple refracting telescope. The objective lens produces an inverted image which is viewed by the eyepiece acting as a magnifier. Note that the parallel light from the star focuses at the focal point of the objective lens. With the image at the focal point of the eyepiece lens, light from the image emerges as parallel rays that are easily focused by the eye.

## Figure 56c

Relationship between the angles  $\theta_0, \theta_i$ , and the focal lengths.

The same formula also applies to a reflecting telescope with  $f_0$  the focal length of the parabolic mirror. Note that there is no arbitrary number like 25 cm in the formula for the magnification of a telescope because telescopes are designed to look at distant objects where the angle  $\theta_0$  the object subtends to the unaided eye is the same for everyone.

The first and the last of the important refracting telescopes are shown in Figures (57). The telescope was invented in Holland in 1608 by Hans Lippershy. Shortly after that, Galileo constructed a more powerful instrument and was the first to use it effectively in astronomy. With a telescope like the one shown in Figure (57a), he discovered the moons of Jupiter, a result that provided an explicit demonstration that heavenly bodies could orbit around something other than the earth. This countered the long held idea that the earth was at the center of everything and provided support for the Copernican sun centered picture of the solar system.

When it comes to building large refracting telescopes, the huge amount of glass in the objective lens becomes a problem. The 1 meter diameter refracting telescope at the Yerkes Observatory, shown in Figure (57b), is the largest refracting telescope ever constructed. That was built back in 1897. The largest reflecting telescope is the new 10 meter telescope at the Keck Observatory at the summit of the inactive volcano Mauna Kea in Hawaii. Since the area and light gathering power of a telescope is proportional to the area or the square of the diameter of the mirror or objective lens, the 10 meter Keck telescope is 100 times more powerful than the 1 meter Yerkes telescope.



**Figure 57a** Galileo's telescope. With such an instrument Galileo discovered the moons of Jupiter.

## **Exercise 8**

To build your own refracting telescope, you purchase a 3 inch diameter objective lens with a focal length of 50 cm. You want the telescope to have a magnification  $m=25\times$ .

(a) What will be the fnumber of your telescope? (1 inch = 2.54 cm).

(b) What should the focal length of your eyepiece lens be?

(c) How far behind the objective lens should the eyepiece lens be located?

(d) Someone give you an eyepiece with a focal length of 10 mm. Using this eyepiece, what magnification do you get with your telescope?

(e) You notice that your new eyepiece is not in focus at the same place as your old eyepiece. Did you have to move the new eyepiece toward or away from the objective lens, and by how much?

(f) Still later, you decide to take pictures with your telescope. To do this you replace the eyepiece with a film holder. Where do you place the film, and why did you remove the eyepiece?



Figure 57b

The Yerkes telescope is the world's largest refracting telescope, was finished in 1897. Since then all larger telescopes have been reflectors.

# **Reflecting telescopes**

In several ways, the reflecting telescope is similar to the refracting telescope. As we saw back in our discussion of parabolic mirrors, the mirror produces an image in the focal plane when the light comes from a distant object. This is shown in Figure (58a) which is similar to our old Figure (4). If you want to look at the image with an eyepiece, you have the problem that the image is in front of the mirror where, for a small telescope, your head would block the light coming into the scope. Issac Newton, who invented the reflecting telescope, solved that problem by placing a small, flat, 45° reflecting surface inside the telescope tube to deflect the image outside the tube as shown in Figure (58b). There the image can easily be viewed using an eyepiece. Newton's own telescope is shown in Figure (58d). Another technique, used in larger telescopes, is to reflect the beam back through a hole in the mirror as shown in Figure (58c).

The reason Newton invented the reflecting telescope was to avoid an effect called *chromatic aberration*. When white light passes through a simple lens, different wavelengths or colors focus at different distances behind the lens. For example if the yellow light is in focus the red and blue images will be out of focus. In contrast, all wavelengths focus at the same point using a parabolic mirror.



**Figure 58d** *Issac Newton's reflecting telescope.* 

However, problems with keeping the reflecting surface shinny, and the development of lens combinations that eliminated chromatic aberration, made refracting telescopes more popular until the late 1800's. The invention of the durable silver and aluminum coatings on glass brought reflecting telescopes into prominence in the twentieth century.



## Figure 58a

A parabolic reflector focuses the parallel rays from a distant object, forming an image a distance  $f_0$  in front of the mirror.



# Figure 58b

Issac Newton's solution to viewing the image was to deflect the beam using a  $45^{\circ}$  reflecting surface so that the eyepiece could be outside the telescope tube.



## Figure 58c

For large telescopes, it is common to reflect the beam back through a hole in the center of the primary mirror. This arrangement is known as the Cassegrain design.

# Large Reflecting Telescopes.

The first person to build a really large reflecting telescope was William Hershel, who started with a two inch reflector in 1774 and by 1789 had constructed the four foot diameter telescope shown in Figure (59a). Among Hershel's accomplishments was the discovery of the planet Uranus, and the first observation a distant nebula. It would be another 130 years before Edwin Hubble, using the 100 inch telescope on Mt. Wilson would conclusively demonstrate that such nebula were in fact galaxies like our own milky way. This also led hubble to discover the expansion of the universe.

During most of the second half of the twentieth century, the largest telescope has been the 200 inch (5 meter) telescope on Mt. Palomar, shown in Figure (59b). This was the first telescope large enough that a person could work at the prime focus, without using a secondary mirror. Hubbel himself is seen in the observing cage at the prime focus in Figure (59c).

Recently it has become possible to construct mirrors larger than 5 meters in diameter. One of the tricks is to cast the molten glass in a rotating container and keep the container rotating while the glass cools. A rotating liquid has a parabolic surface. The faster the rotation the deeper the parabola. Thus by choosing the right rotation speed, one can cast a mirror blank that has the correct parabola built in. The surface is still a bit rough, and has to be polished smooth, but the grinding out oh large amounts of glass is avoided. The 6.5 meter mirror, shown in Figures (59d and e), being installed on top of Mt. Hopkins in Arizona, was built this way. Seventeen tons of glass would have to have been ground out if the parabola had not been cast into the mirror blank.



**Figure 59a** William Hershel's 4 ft diameter, 40 ft long reflecting telescope which he completed in 1789.



Figures 59b,c The Mt. Palomar 200 inch telescope. Below is Edwin Hubble in the observing cage.





Figures 59d,e The 6.5 meter MMT telescope atop Mt. Hopkins. Above, the mirror has not been silvered yet. The blue is a temporary protective coating. Below, the mirror is being hoisted into the telescope frame.



# **Hubbel Space Telescope**

An important limit to telescopes on earth, in their ability to distinguish fine detail, is turbulence in the atmosphere. Blobs of air above the telescope move around causing the star image to move, blurring the picture. This motion, on a time scale of about 1/60 second, is what causes stars to appear to twinkle.

The effects of turbulance, and any distortion caused by the atmosphere, are eliminated by placing the telescope in orbit above the atmosphere. The largest telescope in orbit is the famous Hubble telescope with its 1.5 meter diameter mirror, seen in Figure (60). After initial problems with its optics were fixed, the Hubble telescope has produced fantastic images like that of the Eagle nebula seen in Figure (7-17) reproduced here.

With a modern telescope like the Keck (see next page), the effects of atmospheric turbulance can mostly be eliminated by having a computer can track the image of a bright star. The telescopes mirror is flexible enough that the shape of the mirror can then be be modified rapidly and by a tiny amount to keep the image steady.



**Figure 7-17** The eagle nebula, birthplace of stars. This Hubble photograph, which apeared on the cover of Time magazine, is perhaps the most famous.



**Figure 60a** *The Hubble telescope mirror. How is that for a shaving mirror?* 



**Figure 60b** *Hubble telescope before launch.* 



**Figure 60c** *Hubble telescope being deployed.* 

# **World's Largest Optical Telescope**

As of 1999, the largest optical telescope in the world is the Keck telescope located atop the Mauna Kea volcano in Hawaii, seen in Figure (61a). Actually there are two identical Keck telescopes as seen in the close-up, Figure (61b). The primary mirror in each telescope consists of 36 hexagonal mirrors fitted together as seen in Figure (61c) to form a mirror 10 meters in diameter. This is twice the diameter of the Mt. palomar mirrorwe discussed earlier.

The reason for building two Keck telescopes has to do with the wave nature of light. As we mentioned in the introduction to this chapter, geometrical optics works well when the objects we are studying are large compared to the wavelength of light. This is illustrated by the ripple tank photographs of Figures (33-3) and (33-8) reproduced here. In the left hand figure, we see we see a wave passing through a gap that is considerably wider than the wave's wavelength. On the other side of the gap there is a well defined beam with a distinct shadow. This is what we assume light waves do in geometrical optics.

In contrast, when the water waves encounter a gap whose width is comparable to a wavelength, as in the right hand figure, the waves spread out on the far side. This is a phenomenon called *diffraction*. We can even see some diffraction at the edges of the beam emerging from the wide gap.



Figures 61 a,b The Keck telescopes atop Mauna Kea volcano in Hawaii

Diffraction also affects the ability of telescopes to form sharp images. The bigger the diameter of the telescope, compared to the light wavelength, the less important diffraction is and the sharper the image that can be formed. By combining the output from the two Keck telescopes, one creates a telescope whose effective diameter, for handling diffraction effects, is equal to the 90 meter separation of the telescopes rather than just the 10 meter diameter of one telescope. The great improvement in the image sharpness that results is seen in Figure (61d). On the left is the best possible image of a star, taken using one telescope alone. When the two telescopes are combined, they get the much sharper image on the right.



## Figures 61 c

The 36 mirrors forming Keck's primary mirror. We have emphasized the outline of the upper 4 mirrors.

Figures 33-3,8 Unless the gap is wide in comparison to a wavelength, diffraction effects are important.





**Figures 62** Same star, photographed on the left using one scope, on the right with the two Keck telescopes combined.

# **Infrared Telescopes**

Among the spectacular images in astronomy are the large dust clouds like the ones that form the Eagle nebula photographed by the Hubble telescope, and the famous Horsehead nebula shown in Figure (63a). But a problem is that astronomers would like to see through the dust, to see what is going on inside the clouds and what lies beyond.

While visible light is blocked by the dust, other wavelength's of electromagnetic radiation can penetrate these clouds. Figure (63b) is a photograph of the same patch of sky as the Horsehead nebula in (63a), but observed using infrared light whose wavelengths are about 3 times longer than the wavelengths of visible light. First notice that the brightest stars are at the same positions in both photographs. But then notice that the black cloud, thought to resemble a horses head, is missing in the infrared photograph. The stars in and behind the cloud shine through; their infrared light is not blocked by the dust.

**a)** Visible light photograph





## Figure 63

b)

Infrared light photograph

The horsehead nebula photographed in visible (a) and infrared light (b). The infrared light passes through the dust cloud. Where does the infrared light come from? If you have studied Chapter 35 on the Bohr theory of hydrogen, you will recall that hydrogen atoms can radiate many different wavelengths of light. The only visible wavelengths are the three longest wavelengths in the Balmer series. The rest of the Balmer series and all of the Lyman series consist of short wavelength ultraviolet light. But all the other wavelengths radiated by hydrogen are infrared, like the Paschen series where the electron ends up in the third energy level. The infrared wavelengths are longer than those of visible light. Since hydrogen is the major constituent of almost all stars, it should not be surprising that stars radiate infrared as well as visible light.

A telescope designed for looking at infrared light is essentially the same as a visible light telescope, except for the camera. Figure (64) shows the infrared telescope on Mt. Hopkins used to take the infrared image of the Horsehead nebula. We enlarged the interior photograph to show the infrared camera which is cooled by a jacket of liquid nitrogen (essentially a large thermos bottle surrounding the camera).



## Figure 64

Infrared telescope on Mt. Hopkins. Note that the infrared camera, seen in the blowup, is in a container cooled by liquid nitrogen. You do not want the walls of the camera to be "infrared hot" which would fog the image.



You might wonder why you have to cool an infrared camera and not a visible light camera. The answer is that warm bodies emit infrared radiation. The hotter the object, the shorter the wavelength of the radiation. If an object is hot enough, it begins to glow in visible light, and we say that the object is red hot, or white hot. Since you do not want the infrared detector in the camera seeing camera walls glowing "infrared hot", the camera has to be cooled.

Not all infrared radiation can make it down through the earth's atmosphere. Water vapor, for example is very good at absorbing certain infrared wavelengths. To observe the wavelengths that do not make it through, infrared telescopes have been placed in orbit. Figure 65 is an artist's drawing of the Infrared Astronomical Satellite (IRAS) which was used to make the infrared map of the entire sky seen in Figure (66). The map is oriented so that the Milky Way, our own galaxy, lies along the center horizontal plane. In visible light photographs, most of the stars in our own galaxy are obscured by the immense amount of dust in the plane of the galaxy. But in an infrared photograph, the huge concentration of stars in the plane of the galaxy show up clearly.

At the center of our galaxy is a gigantic black hole, with a mass of millions of suns. For a visible light telescope, the galactic center is completely obscured by dust. But the center can be clearly seen in the infrared photograph of Figure (67), taken by the Mt. Hopkins telescope of Figure 64. This is not a single exposure, instead it is a composite of thousands of images in that region of the sky. Three different infrared wavelengths were recorded, and the color photograph was created by displaying the longest wavelength image as red, the middle wavelength as green, and the shortest wavelength as blue. In this photograph, you not only see the intense radiation from the region of the black hole at the center, but also the enormous density of stars at the center of our galaxy. (You do not see radiation from the black hole itself, but from nearby stars that may be in the process of being captured by the black hole.)



**Figure 65** Artist's drawing of the infrared telescope IRAS in orbit.



**Figure 66** Map of the entire sky made by IRAS. The center of the Milky Way is in the center of the map. This is essentially a view of our galaxy seen from the inside.



## **Figure 67** Center of our galaxy, where an enormous black hole resides. Not only is the galactic center rich in stars, but also in dust which prevents viewing this region in visible light.

# **Radio Telescopes**

The earth's atmosphere allows not only visible and some infrared light from stars to pass through, but also radio waves in the wavelength range from a few millimeters to a good fraction of a meter. To study the radio waves emitted by stars and galaxies, a number of *radio telescopes* have been constructed.

For a telescope reflector to produce a sharp image, the surface of the reflector should be smooth and accurate to within about a fifth of a wavelength of the radiation being studied. For example, the surface of a mirror for a visible wavelength telescope should be accurate to within about  $10^{-4}$  millimeters since the wavelength of visible light is centered around  $5 \times 10^{-4}$  millimeters. Radio telescopes that are to work with 5 millimeter wavelength radio waves, need surfaces accurate only to about a millimeter. Telescopes designed to study the important 21 cm wavelength radiation emitted by hydrogen, can have a rougher surface yet. As a result, radio telescopes can use sheet metal or even wire mesh rather than polished glass for the reflecting surface.

This is a good thing, because radio telescopes have to be much bigger than optical telescopes to order to achieve comparable images. The sharpness of an image, due to diffraction effects, is related to the ratio of the reflector diameter to the radiation wavelength. Since the radio wavelengths are at least 10<sup>4</sup> times larger than those for visible light, a radio telescope has to be 10<sup>4</sup> times larger than an optical telescope to achieve the same resolution.

The worlds largest radio telescope dish, shown in Figure (68), is the 305 meter dish at the Arecibo Observatory in Perto Rico. While this dish can see faint objects because of



**Figure 68** Arecibo radio telescope. While the world's largest telescope dish remains fixed in the earth, the focal point can be moved to track a star.

it's enormous size, and has been used to make significant discoveries, it has the resolving ability of an optical telescope about 3 centimeters in diameter, or a good set of binoculars.

As we saw with the Keck telescope, there is a great improvement in resolving power if the images of two or more telescopes are combined. The effective resolving power is related to the separation of the telescopes rather than to the diameter of the individual telescopes. Figure (69) shows the *Very Large Array (VLA)* consisting of twenty seven 25 meter diameter radio telescopes located in southern New Mexico. The dishes are mounted on tracks, and can be spread out to cover an area 36 kilometers in diameter. At this spacing, the resolving power is nearly comparable to a 5 meter optical telescope at Mt. Palomar.



## Figures 69

The "Very Large Array" (VLA) of radio telescopes. The twenty seven telescopes can be spread out to a diameter of 36 kilometers.



**Figures 69b** *Radio galaxy image from the VLA. Studying the radio waves emitted by a galaxy often gives a very different picture than visible light.* 



## Figure 70

The Very Long Baseline Array of radio antennas. They are located at a) Hancock New Hampshire b) Ft. Davis Texas c) Kitt Peak Arizona d) North Liberty Iowa e) St. Croix Virgin Islands f) Brewster Washington g) Mauna Kea Hawaii h) Pie Town New Mexico i) Los Alamos New Mexico j) Owen's Valley California.

## The Very Long Baseline Array (VLBA)

To obtain significantly greater resolving power, the *Very Long Baseline Array (VLBA)* was set up in the early 1990's. It consists of ten 25 meter diameter radio telescopes placed around the earth as shown in Figure (70). When the images of these telescopes are combined, the resolving power is comparable to an optical telescope 1000 meters in diameter (or an array of optical telescopes spread over an area one kilometer across).

The data from each telescope is recorded on a high speed digital tape with a time track created by a hydrogen maser atomic clock. The tapes are brought to a single location in Socorrow New Mexico where a high speed computer uses the accurate time tracks to combine the data from all the telescopes into a single image. To do this, the computer has to correct, for example, for the time difference of the arrival of the radio waves at the different telescope locations.

Because of it's high resolution, the VLBA can be used to study the structure of individual stars. In Figure 72 we see two time snapshots of the radio emission from the stellar atmosphere of a star 1000 light years away. With any of the current optical telescopes, the image of this star is only a point.



## Figure 71

Very Long Baseline Array (VLBA) radio images of the variable star TX Cam which is located 1000 light years away. The approximate size of the star as it would be seen in visible light is indicated by the circle. The spots are silicon Monoxide (SiO) gas in the star's extended atmosphere. Motion of the these spots trace the periodic changes in the atmosphere of the star. (Credit P.J. Diamond & A.J. Kembal, National Radio Astronomy, Associated Universities, Inc.)

# MICROSCOPES

Optically, microscopes like the one seen in Figure (72), are telescopes designed to focus on nearby objects. Figure (73) shows the ray diagram for a simple microscope, where the objective lens forms an inverted image which is viewed by an eyepiece.

To calculate the magnification of a simple microscope, note that if an object of height  $y_0$  were viewed unaided at a distance of 25 cm, it would subtend an angle  $\theta_0$  given by

$$\theta_0 = \frac{y_0}{25 \text{ cm}} \tag{42}$$

where throughout this discussion we will use the small angle approximation  $\sin\theta \approx \tan\theta \approx \theta$ .

A ray from the tip of the object (point A in Figure 73b), parallel to the axis, will cross the axis at point D, the focal point of the objective lens. Thus the height BC is equal to the height  $y_0$  of the object, and the distance BD is the focal length  $f_0$  of the objective, and the angle  $\beta$ is given by

$$\beta = \frac{y_0}{f_0} \qquad \begin{array}{c} \textit{from triangle} \\ \textit{BCD} \end{array} \tag{43}$$

From triangle DEF, where the small angle at D is also  $\beta$ , we have

$$\beta = \frac{y_i}{L} \qquad from triangle \\ DEF \qquad (44)$$

where  $y_i$  is the height of the image and the distance L is called the *tube length* of the microscope.



Equating the values of  $\beta\,$  in Equations 29 and 30 and solving for  $y_i\,$  gives

$$\beta = \frac{y_0}{f_0} = \frac{y_i}{L} ; \quad y_i = y_0 \frac{L}{f_0}$$
(45)

The eyepiece is placed so that the image of the objective is in the focal plane of the eyepiece lens, producing parallel rays that the eye can focus. Thus the distance EG equals the focal length  $f_0$  of the eyepiece. From triangle EFG we find that the angle  $\theta_i$  that image subtends as seen by the eye is

$$\theta_{i} = \frac{y_{i}}{f_{e}} \qquad angle subtended \\ by image \qquad (46)$$

Substituting Equation 45 for  $y_i$  in Equation 46 gives

$$\theta_{i} = \frac{L}{f_{0}} \frac{y_{0}}{f_{e}}$$
(47)

Finally, the magnification *m* of the microscope is equal to the ratio of the angle  $\theta_i$  subtended by the image in the microscope, to the angle  $\theta_0$  the object subtends at a distance of 25 cm from the unaided eye.

$$m = \frac{\theta_{\rm i}}{\theta_0} = \frac{L}{f_0} \frac{y_0}{f_{\rm e}} \times \frac{1}{y_0/25 \,{\rm cm}}$$
 (48)

where we used Equation 47 for  $\theta_i$  and Equation 42 for  $\theta_0$ . The distance  $y_0$  cancels in Equation 48 and we get

$$m = \frac{L}{f_0} \times \frac{25 \text{ cm}}{f_e} \qquad \qquad \begin{array}{c} magnification of a\\ simple \ microscope \end{array}$$
(49)

(We could have inserted a minus sign in the formula for magnification to indicate that the image is inverted.)



Figure 72

Standard optical microscope, which my grandfather purchased as a medical student in the 1890's. Compare this with a microscope constructed 100 years later, seen in Figure (69) on the next page.

Optics of a simple microscope.

# **Scanning Tunneling Microscope**

Modern research microscopes bear less resemblance to the simple microscope described above than the Hubble telescope does to Newton's first reflector telescope. In the research microscopes that can view and manipulate individual atoms, there are no lenses based on geometrical optics. Instead the surface to be studied is scanned,



## a) Probe and sample holder.

b) Vacuum chamber enclosing the probe and sample holder. Photograph taken in Geoff Nunes' lab at Dartmouth College.

## Figure 74

Scanning Tunneling Microscope (STM). The tungsten probe seen in (a) has a very sharp point, about one atom across. With a couple of volts difference between the probe and the silicon crystal in the sample holder, an electric current begins to flow when the tip gets to within about fifteen angstroms (less than fifteen atomic diameters) of the surface. The current flows because the wave nature of the electrons allows them to "tunnel" through the few angstrom gap. The current increases rapidly as the probe is brought still closer. By moving the probe in a line sideways across the face of the silicon, while moving the probe in and out to keep the current constant, the tip of the probe travels at a constant height above the silicon atoms. By recording how much the probe was moved in and out, one gets a recording of the shape of the surface along that line. By scanning across many closely spaced lines, one gets a map of the entire surface. The fine motions of the tungsten probe are controlled by piezo crystals which expand or contract by tiny amounts when a voltage is applied to them. The final image you see was created by computer from the scanning data.

<image>

line by line, by a tiny probe whose operation is based on the particle-wave nature of electrons. An image of the

surface is then reconstructed by computer and displayed

on a computer screen. These microscopes work at a

scale of distance much smaller than the wavelength of



c) Surface (111 plane) of a silicon crystal imaged by this microscope. We see the individual silicon atoms in the surface

## PHOTOGRAPH CREDITS

Figure 36-1, p1, p8; Scattered Wave Education Development Center Figure 33-30, p10; Shock wave **Education Development Center** Figures Optics-1 p3; Mormon Tabrenacle The Church of Jesus Christ of Latter-day Saints, Historical Department Archives Figure Optics-8b p7; Corner reflectors NASA Figure Optics-9, p9; Wave through lens Nils Abramson Figure Optics-10, p11; Refraction, ripple tank Education Development Center Figure Optics-11, p9; Refraction, glass ©1990 Richard Megna/ Fundamental Photographs Figure Optics-16b, *p14*; Glass Fiber Foto Forum Figure Optics-18, p15; Duodenum Dr. Richard Rothstein Figure Optics-19, p18; Halo © Robert Greenler Figure Optics-25, p18; Zoom lens Nikon Figure Optics-28, p22; Hubble Scope NASA Figure Optics-42, p31; Human eye © Lennart Nilsson Figure Optics-43, p31; Rods & cones © Lennart Nilsson Figure Optics-47b, p33; Single lens reflex Adapted from Nikon drawing Figure Optics-57a, p41; Galileo's telescope © Institute and Museum of History of Science of Florence Italy Figure Optics-57b, p41; Yerkes telescope University of Chicago

Figure Optics-58d, p42; Newton's telescope Dorling Kindersley, Pockets Inventions Figure Optics-59a, p43; Hershel's telescope © Royal Astronomical Society Library Figure Optics-59b.c, p43; Palomar telescope Courtesy of The Archives, California Institute of Technology Figure Optics-59d,e, p43; MMT telescope Lori Stiles, Universitu of Arizona News Service Figure 7-17, p44; Eagle nebula Space Telescope Science Institute/NASA Figure Optics-60, p44; Hubble telescope NASA Figure Optics-61a,b,c, p45; Keck telescope © Richard Wainscoat Figure 33-3,8, *p*45; Wave through gap Education Development Center Figure Optics-62, p45; Sharpened star image Keck Figure Optics-63, p46; Horsehead nebula Two Micron All Sky Survay (2MASS) Figure Optics-64, p46; Infrared telescope Two Micron All Sky Survay (2MASS) Figure Optics-65, p47; IRAS drawing Space Infrared Telescope Facility Figure Optics-66, p47; Iras Milky Way Space Infrared Telescope Facility Figure Optics-67, p47; Center of galaxy Two Micron All Sky Survay (2MASS) Figure Optics-68, p48; Arecibo telescope National Astronomy and Ionosphere Center Figure Optics-69a, p48; VLA radio telescopes Photo courtesy of NRAO/AUI. Figure Optics-69b, p48; Radio galaxy Photo courtesy of NRAO/AUI. Figure Optics-70, p49; Radio image of star

Photo courtesy of NRAO/AUI.

# Calculus 2000 A Physics Based Calculus Text

The Physics 2000 (P2000) text uses certain calculus concepts that are taught in Chapter 1 of this Calculus 2000 (C2000) text. For students whose calculus is rusty, or who have not had calculus, they should start studying Chapter 1 of C2000 while studying Chapter 4 of P2000 on the use of calculus in Physics. By the time the student reaches Part 2 of P2000, Chapter 23, the ideas contained in Chapter 1 of C2000 should be well understood. That is because the chapters on electric and magnetic fields make extensive use of these basic calculus ideas.

The remaining chapters in Calculus 2000 start from physical concepts built up in P2000 and introduce the student to advanced mathematical techniques. These include concepts such as gradient, divergence, and curl which are essential tools for further study of physics and engineering. Later chapters in C2000 will include such topics as an introduction to complex variables, the Lorentz transformation and 4 vector notation, and two chapters on fluid dynamics.

In the standard calculus text, there is an advantage to presenting a theorem in its most general form so that the theorem can be used effectively in later proofs. The emphasis is on the logical structure of the mathematics. The problem the physics student often encounters is that before the intuitive implications of one theorem can be worked out, the instructor has plowed through five more theorems and proofs. The mathematical structure may be clear, but the usefulness of the mathematics is obscured.

In Calculus 2000, the emphasis is on the intuitive use of the mathematics. We do not introduce a new mathematical concept or technique until the foundation has been developed in the Physics 2000 text. In Chapter 3 of P2000, for example, we use strobe photographs to introduce the concepts of velocity, acceleration, and the limiting process. The calculus limit at  $\Delta t$ goes to zero is represented physically by the idea of turning the strobe flashing rate all the way up. We point out, however, that because of the uncertainty principle we reach a point where further increase of the flashing rate affects the behavior of the object being studied. Before the student works with calculus formulas, we set both the intuitive basis for the calculus concepts and discuss the limitation of their applicability.

As we mentioned, Calculus 2000 provides the calculus background necessary to complete all of the Physics 2000 text. All further calculus and advanced mathematics concepts used in P2000 are developed in the physics text. Chapter 2 of P2000 discusses vectors and their scalar and vector products. The scalar dot product is applied in Chapter 8 to the discussion of work and energy, and the vector cross product is applied in the discussion of torques and gyroscopes in Chapter 12. In Chapter 14 on oscillations, we introduce the student to some techniques of solving differential equations. The idea is to guess an answer and plug the guess into the equation to see if it works. The important point, however, is that we use experimental results to make an informed guess. In Chapter 16 the student is introduced to Fourier analysis. We wrote the MacScope<sup>TM</sup> computer program so that the student could easily use Fourier analysis to study experimental data. One of our standard laboratory examples is to use Fourier analysis to determine the normal modes of oscillation of a system of coupled air carts.

Part II of Physics 2000, particularly Chapters 23 through 32 on Electric and Magnetic fields, are more mathematically based than the other chapters. Chapter 23 on fluids is used to introduce the concept of a vector field. It is easier to visualize the velocity field of a fluid with its streamlines, than the more abstract electric and magnetic field lines. We also use the velocity field to introduce the concepts involved in Gauss' law.

In Chapter 25 we use contour maps to introduce equipotential lines and the concept of electrical voltage. Chapter 29 formalizes Gauss' law as an example of a surface integral, and introduces the closed line integral in the discussion of Ampere's law. By pointing out that equations for both the surface integral and the line integral are required to uniquely determine a vector field, the student sees why four equations are needed to specify both the electric field and the magnetic field. These are the four integral equations that form Maxwell's equations studied in Chapter 32. Our main mathematical achievement in Chapter 32 is to show that a pulse of crossed electric and magnetic fields travels through space at the speed of light.

Chapters 33 through 40 of P2000 are not as mathematically focused because we concentrate on developing an intuitive picture of the particle/wave nature of matter. Because we have used special relativity throughout the text, it is easy to introduce the zero rest mass photons as the particle nature of light waves. The wave nature of the electron is introduced with de Broglie's hypothesis and an electron diffraction experiment that is similar to the laser diffraction experiments of Chapter 33. Fourier analysis plays a basic role in showing how the particle/ wave nature of matter leads to the uncertainty principle discussed in Chapter 40. After Chapter 1, the remaining chapters in Calculus 2000 use physical concepts developed in P2000 to introduce advanced calculus techniques needed by physicists and engineers. The concept of a second derivative, and the differential form of the wave equation follows from our discussion of one dimensional waves in Chapter 12 of P2000. The student sees how much easier it is to determine the speed of a wave from the differential wave equation than it is to go through the non calculus arguments we used in Chapter 15 of the P2000 text.

Once the student is familiar with the field plotting models and mapping techniques discussed in Chapter 25 of P2000, she or he is ready for the discussion of the gradient function  $\vec{\nabla} \phi$  described in Chapter 3 of C2000. In Chapter 4 of C2000 the divergence and curl operation are presented as the differential form of the surface and line integrals discussed in Chapter 29 of P2000. Maxwell's equations discussed in Chapter 32 of P2000 are presented in differential form in Chapter 5 of C2000, and from them we derive the differential form of the wave equation for electric and magnetic fields. Again we see that it is easier to determine the speed of a wave from the differential wave equation than from the fairly complex derivation we carried out using the integral equation. Chapter 5 ends with the introduction of the vector potential to simplify the wave equation when source terms are involved.





# Figure 1

Transition to instantaneous velocity.

# Calculus Chapter 1 Introduction to Calculus

This chapter, which replaces Chapter 4 in Physics 2000, is intended for students who have not had calculus, or as a calculus review for those whose calculus is not well remembered. If, after reading part way through this chapter, you feel your calculus background is not so bad after all, go back to Chapter 4 in Physics 2000, study the derivation of the constant acceleration formulas beginning on page 4-8, and work the projectile motion problems in the appendix to Chapter 4. Those who study all of this introduction to calculus should then proceed to the projectile motion problems in the appendix to Chapter 4 of the physics text.

In Chapter 3 of Physics 2000, we used strobe photographs to define velocity and acceleration vectors. The basic approach was to turn up the strobe flashing rate as we did in going from Figure (3-3) to (3-4) until all the kinks are clearly visible and the successive displacement vectors give a reasonable description of the motion. We did not turn the flashing rate too high, for the practical reason that the displacement vectors became too short for accurate work.

# LIMITING PROCESS

In our discussion of instantaneous velocity we conceptually turned the strobe all the way up as illustrated in Figures (2-22a) through (2-22d), redrawn here in Figure (1). In these figures, we initially see a fairly large change in  $\vec{v}_0$  as the strobe rate is increased and  $\Delta t$ reduced. But the change becomes smaller and it looks as if we are approaching some final value of  $\vec{v}_0$  that does not depend on the size of  $\Delta t$ , provided  $\Delta t$  is small enough. It looks as if we have come close to the final value in Figure (1c).

The progression seen in Figure (1) is called a *limiting process*. The idea is that there really is some true value of  $\vec{v}_0$  which we have called the instantaneous velocity, and that we approach this true value for sufficiently small values of  $\Delta t$ . This is a calculus concept, and in the language of calculus, we are *taking the limit as*  $\Delta t$  *goes to zero*.

## **The Uncertainty Principle**

For over 200 years, from the invention of calculus by Newton and Leibnitz until 1924, the limiting process and the resulting concept of instantaneous velocity was one of the cornerstones of physics. Then in 1924 Werner Heisenberg discovered what he called the *uncertainty principle* which places a limit on the accuracy of experimental measurements. Heisenberg discovered something very new and unexpected. He found that the act of making an experimental measurement unavoidably affects the results of an experiment. This had not been known previously because the effect on large objects like golf balls is undetectable. But on an atomic scale where we study small systems like electrons moving inside an atom, the effect is not only observable, it can dominate our study of the system.

One particular consequence of the uncertainly principle is that the more accurately we measure the position of an object, the more we disturb the motion of the object. This has an immediate impact on the concept of instantaneous velocity. If we turn the strobe all the way up, reduce  $\Delta t$  to zero, we are in effect trying to measure the position of the object with infinite precision. The consequence would be an infinitely big disturbance of the motion of the object we are studying. If we actually could turn the strobe all the way up, we would destroy the object we were trying to study.

It turns out that the uncertainty principle can have a significant impact on a larger scale of distance than the atomic scale. Suppose, for example, that we constructed a chamber 1 cm on a side, and wished to study the projectile motion of an electron inside. Using Galileo's idea that objects of different mass fall at the same rate, we would expect that the motion of the electron projectile should be the same as more massive objects. If we took a strobe photograph of the electron's motion, we would expect get results like those shown in Figure (2). This figure represents projectile motion with an acceleration g = 980 cm/sec<sup>2</sup> and  $\Delta t = .01$  sec, as the reader can easily check.



When we study the uncertainty principle in Chapter 40 of the physics text, we will see that a measurement which is accurate enough to show that Position (2) is below Position (1), could disturb the electron enough to reverse its direction of motion. The next position measurement could find the electron over where we drew Position (3), or back where we drew Position (0), or anywhere in the region in between. As a result we could not even determine what direction the electron is moving. This uncertainty would not be the result of a sloppy experiment, it is the best we can do with the most accurate and delicate measurements possible.

The uncertainty principle has had a significant impact on the way physicists think about motion. Because we now know that the measuring process affects the results of the measurement, we see that it is essential to provide experimental definitions to any physical quantity we wish to study. A conceptual definition, like turning the strobe all the way up to define instantaneous velocity, can lead to fundamental inconsistencies.

Even an experimental definition like our strobe definition of velocity can lead to inconsistent results when applied to something like the electron in Figure (2). But these inconsistencies are real. Their existence is telling us that the very concept of velocity is beginning to lose meaning for these small objects.

On the other hand, the idea of the limiting process and instantaneous velocity is very convenient when applied to larger objects where the effects of the uncertainty principle are not detectable. In this case we can apply all the mathematical tools of calculus developed over the past 250 years. The status of instantaneous velocity has changed from a basic concept to a useful mathematical tool. Those problems for which this mathematical tool works are called problems in *classical physics*; and those problems for which the uncertainty principle is important, are in the realm of what we call *quantum physics*.

**Figure 2** *Hypothetical electron projectile motion experiment.* 

## CALCULUS DEFINITION OF VELOCITY

With the above perspective on the physical limitations on the limiting process, we can now return to the main topic of this chapter—the use of calculus in defining and working with velocity and acceleration.

In discussing the limiting process in calculus, one traditionally uses a special set of symbols which we can understand if we adopt the notation shown in Figure (3). In that figure we have drawn the coordinate vectors  $\vec{R}_i$  and  $\vec{R}_{i+1}$  for the i th and (i + 1) th positions of the object. We are now using the symbol  $\vec{\Delta R}_i$  to represent the displacement of the ball during the i to i+1 interval. The vector equation for  $\vec{\Delta R}_i$  is

$$\vec{\Delta R}_i = \vec{R}_{i+1} - \vec{R}_i \tag{1}$$

In words, Equation (1) tells us that  $\overrightarrow{\Delta R}_i$  is the change, during the time  $\Delta t$ , of the position vector  $\vec{R}$  describing the location of the ball.



**Figure 3** Definitions of  $\overline{\Delta R}_i$  and  $\vec{v}_i$ .

The velocity vector  $\vec{v}_i$  is now given by

$$\vec{v}_i = \frac{\vec{\Delta R}_i}{\Delta t}$$
(2)

This is just our old strobe definition  $\vec{v}_i = \vec{s}_i/\Delta t$ , but using a notation which emphasizes that the displacement  $\vec{s}_i = \Delta \vec{R}_i$  is the *change in position* that occurs during the time  $\Delta t$ . The Greek letter  $\Delta$  (delta) is used both to represent the idea that the quantity  $\Delta \vec{R}_i$  or  $\Delta t$ is small, and to emphasize that both of these quantities change as we change the strobe rate.

The limiting process in Figure (1) can be written in the form

$$\vec{\mathbf{y}}_{i} \equiv \lim_{\Delta t \to 0} \frac{\overline{\Delta R}_{i}}{\Delta t}$$
(3)

where the word "limit" with  $\Delta t \rightarrow 0$  underneath, is to be read as "limit as  $\Delta t$  goes to zero". For example we would read Equation (3) as "*the instantaneous velocity*  $\vec{y}_i$  *at position* i *is the limit, as*  $\Delta t$  *goes to zero, of the ratio*  $\Delta \vec{R}_i / \Delta t$ ."

For two reasons, Equation (3) is not quite yet in standard calculus notation. One is that in calculus, only the limiting value, in this case, the instantaneous velocity, is considered to be important. Our strobe definition  $\vec{v}_i = \vec{\Delta R}_i / \Delta t$  is only a step in the limiting process. Therefore when we see the vector  $\vec{v}_i$ , we should assume that it is the limiting value, and no special symbol like the underline is used. For this reason we will drop the underline and write

$$\vec{v}_i = \underset{\Delta t \to 0}{\text{limit}} \frac{\Delta R_i}{\Delta t}$$
 (3a)
The second change deals with the fact that when  $\Delta t$  goes to zero we need an infinite number to time steps to get through our strobe photograph, and thus it is not possible to locate a position by counting time steps. Instead we measure the time t that has elapsed since the beginning of the photograph, and use that time to tell us where we are, as illustrated in Figure (4). Thus instead of using  $\vec{v}_i$  to represent the velocity at position i, we write  $\vec{v}(t)$  to represent the velocity at time t. Equation (3) now becomes

$$\vec{v}(t) = \underset{\Delta t \to 0}{\text{limit}} \frac{\Delta \vec{R}(t)}{\Delta t}$$
 (3b)

where we also replaced  $\overline{\Delta R}_i$  by its value  $\overline{\Delta R}(t)$  at time t.

Although Equation (3b) is in more or less standard calculus notation, the notation is clumsy. It is a pain to keep writing the word "limit" with a  $\Delta t \rightarrow 0$  underneath. To streamline the notation, we replace the Greek letter  $\Delta$  with the English letter d as follows

$$\underset{\Delta t \to 0}{\text{limit}} \frac{\overrightarrow{\Delta R}(t)}{\Delta t} \equiv \frac{d\vec{R}(t)}{dt}$$
(4)

(The symbol  $\equiv$  means *defined equal to*.) To a mathematician, the symbol  $d\vec{R}(t)/dt$  is just shorthand



#### Figure 4

Rather than counting individual images, we can locate a position by measuring the elapsed time t. In this figure, we have drawn the displacement vector  $\vec{R}(t)$  at time t = .3 sec. notation for the limiting process we have been describing. But to a physicist, there is a different, more practical meaning. Think of dt as a short  $\Delta t$ , short enough so that the limiting process has essentially occurred, but not too short to see what is going on. In Figure (1), a value of dt less than .025 seconds is probably good enough.

If dt is small but finite, then we know exactly what the  $d\vec{R}(t)$  is. It is the small but finite displacement vector at the time t. It is our old strobe definition of velocity, with the added condition that dt is such a short time interval that the limiting process has occurred. From this point of view, dt is a real time interval and  $d\vec{R}(t)$  a real vector, which we can work with in a normal way. The only thing special about these quantities is that when we see the letter d instead of  $\Delta$ , we must remember that a limiting process is involved. In this notation, the calculus definition of velocity is

$$\vec{v}(t) = \frac{d\vec{R}(t)}{dt}$$
(5)

where  $\vec{R}(t)$  and  $\vec{v}(t)$  are the particle's coordinate vector and velocity vector respectively, as shown in Figure (5). Remember that this is just fancy shorthand notation for the limiting process we have been describing.



**Figure 5** *Instantaneous position and velocity at time t.* 

# ACCELERATION

In the analysis of strobe photographs, we defined both a velocity vector  $\vec{v}$  and an acceleration vector  $\vec{a}$ . The definition of  $\vec{a}$ , shown in Figure (2-12) reproduced here in Figure (6) was

$$\vec{a}_{i} \equiv \frac{\vec{v}_{i+1} - \vec{v}_{i}}{\Delta t} \tag{6}$$

In our graphical work we replaced  $\vec{v}_i$  by  $\vec{s}_i/\Delta t$  so that we could work directly with the displacement vectors  $\vec{s}_i$  and experimentally determine the behavior of the acceleration vector for several kinds of motion.

Let us now change this graphical definition of acceleration over to a calculus definition, using the ideas just applied to the velocity vector. First, assume that the ball reached position i at time t as shown in Figure (6). Then we can write

$$\vec{v}_i = \vec{v}(t)$$
  
 $\vec{v}_{i+1} = \vec{v}(t+\Delta t)$ 

to change the time dependence from a count of strobe flashes to the continuous variable t. Next, define the vector  $\Delta \vec{v}(t)$  by

$$\vec{\Delta v}(t) \equiv \vec{v}(t+\Delta t) - \vec{v}(t) \quad \left( = \vec{v}_{i+1} - \vec{v}_i \right)$$
(7)

We see that  $\overrightarrow{\Delta v}(t)$  is the change in the velocity vector as the time advances from t to  $t+\Delta t$ . The strobe definition of  $\vec{a}_i$  can now be written

$$\vec{a}(t) \begin{pmatrix} strobe\\ definition \end{pmatrix} = \frac{\vec{v}(t + \Delta t) - \vec{v}(t)}{\Delta t} \equiv \frac{\Delta \vec{v}(t)}{\Delta t} \quad (8)$$



**Figure 6** *Experimental definition of the acceleration vector.* 

Now go through the limiting process, turning the strobe up, reducing  $\Delta t$  until the value of  $\vec{a}(t)$  settles down to its limiting value. We have

$$\vec{a}(t) \begin{pmatrix} calculus \\ definition \end{pmatrix} = \lim_{\Delta t \to 0} \frac{\vec{v}(t + \Delta t) - \vec{v}(t)}{\Delta t}$$

$$= \lim_{\Delta t \to 0} \frac{\vec{\Delta v}(t)}{\Delta t}$$
(9)

Finally use the shorthand notation d/dt for the limiting process:

$$\vec{a}(t) = \frac{d\vec{v}(t)}{dt}$$
(10)

Equation (10) does not make sense unless you remember that it is notation for all the ideas expressed above. Again, physicists think of dt as a short but finite time interval, and  $d\vec{v}(t)$  as the small but finite change in the velocity vector during the time interval dt. It's our strobe definition of acceleration with the added requirement that  $\Delta t$  is short enough that the limiting process has already occurred.

## Components

Even if you have studied calculus, you may not recall encountering formulas for the derivatives of vectors, like  $d\vec{R}(t)/\Delta t$  and  $d\vec{v}(t)/\Delta t$  which appear in Equations (5) and (10). To bring these equations into a more familiar form where you can apply standard calculus formulas, we will break the vector Equations (5) and (10) down into component equations.

In the chapter on vectors, we saw that any vector equation like

$$\vec{A} = \vec{B} + \vec{C} \tag{11}$$

is equivalent to the three component equations

$$A_x = B_x + C_x$$
  

$$A_y = B_y + C_y$$
  

$$A_z = B_z + C_z$$
(12)

The advantage of the component equations was that they are simply numerical equations and no graphical work or trigonometry is required. The limiting process in calculus does not affect the decomposition of a vector into components, thus Equation (5) for  $\vec{v}(t)$  and Equation (10) for  $\vec{a}(t)$  become

$$\vec{v}(t) = d\vec{R}(t)/dt$$
(5)

$$v_{x}(t) = dR_{x}(t)/dt$$
 (5a)

$$v_{\rm v}(t) = dR_{\rm v}(t)/dt \tag{5b}$$

$$v_z(t) = dR_z(t)/dt$$
 (5c)

and

$$\vec{a}(t) = d\vec{v}(t)/dt \tag{10}$$

$$a_{x}(t) = dv_{x}(t)/dt$$
(10a)

$$a_{y}(t) = dv_{y}(t)/dt$$
(10b)

$$a_{z}(t) = dv_{z}(t)/dt$$
(10c)

Often we use the letter x for the x coordinate of the vector  $\vec{R}$  and we use y for  $R_y$  and z for  $R_z$ . With this notation, Equation (5) assumes the shorter and perhaps more familiar form

$$v_{x}(t) = dx(t)/dt$$

$$v_{y}(t) = dy(t)/dt$$

$$v_{z}(t) = dz(t)/dt$$
Figure 7
(5a')
(5b')
(5b')
(5c')

At this point the notation has become deceptively short. You now have to remember that x(t) stands for the x coordinate of the particle at a time t.

We have finally boiled the notation down to the point where it would be familiar from any calculus course. If we restrict our attention to one dimensional motion along the x axis. Then all we have to concern ourselves with are the x component equations

$$v_{x}(t) = \frac{dx(t)}{dt}$$

$$a_{x}(t) = \frac{dv_{x}(t)}{dt}$$
(10a)

### INTEGRATION

When we worked with strobe photographs, the photograph told us the position  $\vec{R}(t)$  of the ball as time passed. Knowing the position, we can then use Equation (5) to calculate the ball's velocity  $\vec{v}(t)$  and then Equation (10) to determine the acceleration  $\vec{a}(t)$ . In general, however, we want to go the other way, and predict the motion from a knowledge of the acceleration. For example, imagine that you were in Galileo's position, hired by a prince to predict the motion of cannonballs. You know that a cannonball should not be much affected by air resistance, thus the acceleration throughout its trajectory should be the constant gravitational acceleration  $\vec{g}$ . You know that  $\vec{a}(t) = \vec{g}$ ; how then do you use that knowledge in Equations (5) and (10) to predict the motion of the ball?

The answer is that you cannot with the equations in their present form. The equations tell you how to go from  $\vec{R}(t)$  to  $\vec{a}(t)$ , while to predict motion you need to go the other way, from  $\vec{a}(t)$  to  $\vec{R}(t)$ . The topic of this section is to see how to reverse the directions in which we use our calculus equations. Equations (5) and (10) involve the process called *differentiation*. We will see that when we go the other way the reverse of differentiation is a process called *integration*. We will see that integration is a simple concept, but a process that is sometimes hard to perform without the aid of a computer.

## **Prediction of Motion**

In our earlier discussion, we have used strobe photographs to analyze motion. Let us see what we can learn from such a photograph for predicting motion. Figure (8) is our familiar projectile motion photograph showing the displacement  $\vec{s}$  of a ball during the time the ball traveled from a position labeled (0) to the position labeled (4). If the ball is now at position (0) and each of the images is .1 seconds apart, then the vector  $\vec{s}$  tells us where the ball will be at a time of .4 seconds from now. If we can predict  $\vec{s}$ , we can predict the motion of the ball. The general problem of predicting the motion of the ball is to be able to calculate  $\vec{s}(t)$  for any time t.

From Figure (8) we see that  $\vec{s}$  is the vector sum of the individual displacement vectors  $\vec{s}_1$ ,  $\vec{s}_2$ ,  $\vec{s}_3$  and  $\vec{s}_4$ 

$$\vec{s} = \vec{s}_1 + \vec{s}_2 + \vec{s}_3 + \vec{s}_4 \tag{11}$$

We can then use the fact that  $\vec{s}_1 = \vec{v}_1 \Delta t$ ,  $\vec{s}_2 = \vec{v}_2 \Delta t$ , etc. to get

$$\vec{s} = \vec{v}_1 \Delta t + \vec{v}_2 \Delta t + \vec{v}_3 \Delta t + \vec{v}_4 \Delta t$$
(12)

Rather than writing out each term, we can use the *summation sign*  $\Sigma$  to write

$$\vec{s} = \sum_{i=1}^{4} \vec{v}_i \Delta t \tag{12a}$$

Equation (12) is approximate in that the  $\vec{v}_i$  are approximate (strobe) velocities, not the instantaneous veloci-



Figure 8

To predict the total displacement  $\vec{s}$ , we add up the individual displacements  $\vec{s}_i$ .

ties we want for a calculus discussion. In Figure (9) we improved the situation by cutting  $\Delta t$  to 1/4 of its previous value, giving us four times as many images and more accurate velocities  $\vec{v}_i$ .

We see that the displacement  $\vec{s}$  is now the sum of 16 vectors

$$\vec{s} = \vec{s}_1 + \vec{s}_2 + \vec{s}_3 + \dots + \vec{s}_{15} + \vec{s}_{16}$$
 (13)

Expressing this in terms of the velocity vectors  $\vec{v}_1$  to  $\vec{v}_{16}$  we have

$$\vec{s} = \vec{v}_1 \Delta t + \vec{v}_2 \Delta t + \vec{v}_3 \Delta t + ... + \vec{v}_{15} \Delta t + \vec{v}_{16} \Delta t$$
 (14)

or using our more compact notation

$$\vec{s} = \sum_{i=1}^{16} \vec{v}_i \Delta t \tag{14a}$$

While Equation (14) for  $\vec{s}$  looks quite different than Equation (12), the sum of sixteen vectors instead of four, the displacement vectors  $\vec{s}$  in the two cases are exactly the same. Adding more intermediate images did not change where the ball was located at the time of t = .4 seconds. In going from Equation (12) to (14), what has changed in shortening the time step  $\Delta t$ , is that *the individual velocity vectors*  $\vec{v}_i$  *become more nearly equal to the instantaneous velocity of the ball at each image*.



Figure 9 With a shorter time interval, we add up more displacement vectors to get the total displacement  $\vec{s}$ .

If we reduced  $\Delta t$  again by another factor of 1/4, so that we had 64 images in the interval t = 0 to t = .4 sec, the formula for  $\vec{s}$  would become

$$\vec{s} = \sum_{i=1}^{64} \vec{v}_i \Delta t \tag{15a}$$

where now the  $\vec{v}_i$  are still closer to representing the ball's instantaneous velocity. The more we reduce  $\Delta t$ , the more images we include, the closer each  $\vec{v}_i$  comes to the instantaneous velocity  $\vec{v}(t)$ . While adding more images gives us more vectors we have to add up to get the total displacement  $\vec{s}$ , there is very little change in our formula for  $\vec{s}$ . If we had a million images, we would simply write

$$\vec{s} = \sum_{i=1}^{1000000} \vec{v}_i \Delta t$$
 (16a)

In this case the  $\vec{v}_i$  would be physically indistinguishable from the instantaneous velocity  $\vec{v}(t)$ . We have essentially reached a calculus limit, but we have problems with the notation. It is clearly inconvenient to label each  $\vec{v}_i$  and then count the images. Instead we would like notation that involves the instantaneous velocity  $\vec{v}(t)$  and expresses the beginning and end points in terms of the initial time  $t_1$  and final time  $t_f$ , rather than the initial and final image numbers i.

In the calculus notation, we replace the summation sign  $\Sigma$  by something that looks almost like the summation sign, namely the *integral sign*  $\iint$ . (The French word for integration is the same as their word for summation.) Next we replaced the individual  $\vec{v}_i$  by the continuous variable  $\vec{v}(t)$  and finally express the end points by the initial time  $t_i$  and the final time  $t_f$ . The result is

$$\vec{s} = \sum_{i=1}^{n} \vec{v}_{i} \Delta t \rightarrow \begin{pmatrix} as \ the number \\ n \ becomes \\ infinitely \\ large \end{pmatrix} \int_{t_{i}}^{t_{f}} \vec{v}(t) dt \qquad (17)$$

Calculus notation is more easily handled, or is at least more familiar, if we break vector equations up into component equations. Assume that the ball started at position i which has components  $x_i = x(t_i)$  [read  $x(t_i)$ as "x at time  $t_i$ "] and  $y_i = y(t_i)$  as shown in Figure (10). The final position f is at  $x_f = x(t_f)$  and  $y_f = y(t_f)$ . Thus the displacement  $\vec{s}$  has x and y components

$$s_x = x(t_f) - x(t_i)$$
  

$$s_y = y(t_f) - y(t_i)$$

Breaking Equation (17) into component equations gives

$$s_x = x(t_f) - x(t_i) = \int_{t_i}^{t_f} v_x(t) dt$$
 (18a)

$$s_y = y(t_f) - y(t_i) = \int_{t_i}^{t_f} v_y(t) dt$$
 (18b)

Here we will introduce one more piece of notation often used in calculus courses. On the left hand side of Equation (18a) we have  $x(t_f) - x(t_i)$  which we can think of as the variable x(t) evaluated over the interval of time from  $t_i$  to  $t_f$ . We will often deal with variables evaluated over some interval and have a special notation for that. We will write

$$\mathbf{x}(\mathbf{t}_{\mathrm{f}}) - \mathbf{x}(\mathbf{t}_{\mathrm{i}}) \equiv \mathbf{x}(\mathrm{t}) \Big|_{\mathbf{t}_{\mathrm{i}}}^{\mathrm{t}_{\mathrm{f}}}$$
(19)

You are to read the symbol  $x(t) |_{t_i}^{t_f}$  as "x of t evaluated from  $t_i$  to  $t_f$ ". We write the initial time  $t_i$  at the bottom of the vertical bar, the final time  $t_f$  at the top.



Breaking the vector  $\vec{s}$  into components.

We use similar notation for any kind of variable, for example

$$f(x)\Big|_{x_1}^{x_2} \equiv f(x_2) - f(x_1)$$
(19a)

Remember to subtract the variable when evaluated at the value at the bottom of the vertical bar.

With this notation, our Equation (18) can be written

$$s_{x} = x(t) \Big|_{t_{i}}^{t_{f}} = \int_{t_{i}}^{t_{f}} v_{x}(t) dt \qquad (18a')$$

$$s_y = y(t) \Big|_{t_i}^{t_f} = \int_{t_i}^{t_f} v_y(t) dt$$
 (18b')

#### **Calculating Integrals**

Equation (20) is nice and compact, but how do you use it? How do you calculate integrals? The key is to remember that an integral is just a fancy notation for a sum of terms, where we make the time step  $\Delta(t)$  very small. Keeping this in mind, we will see that there is a very easy way to interpret an integral.



Figure 11a

Strobe photograph of ball moving at constant velocity in x direction.





Graph of  $v_{x}(t)$  versus t for the ball of Figure 11a.



**Figure 11c** Each  $v_x \Delta t$  is the area of a rectangle.

To get this interpretation, let us start with the simple case of a ball moving in a straight line, for instance, the x direction, at a constant velocity  $v_x$ . A strobe picture of this motion would look like that shown in Figure (11a).

Figure (11b) is a graph of the ball's velocity  $v_x(t)$  as a function of the time t. The vertical axis is the value of  $v_x$ , the horizontal axis is the time t. Since the ball is traveling at constant velocity,  $v_x$  has a constant value and is thus represented by a straight horizontal line. In order to calculate the distance that the ball has traveled during the time interval from  $t_i$  to  $t_f$ , we need to evaluate the integral

$$s_{x} = \int_{t_{i}}^{t_{f}} v_{x}(t) dt \qquad \begin{array}{c} \text{distance ball} \\ \text{travels in} \\ \text{time interval} \\ t_{i} \text{ to } t_{f} \end{array}$$
(18a)

To actually evaluate the integral, we will go back to our summation notation

$$s_{x} = \sum_{i_{\text{initial}}}^{i_{\text{final}}} v_{xi} \Delta t$$
(20)

and show individual time steps  $\Delta t$  in the graph of  $v_x$  versus t, as in Figure (11c).

We see that each term in Equation (20) is represented in Figure (11c) by a rectangle whose height is  $v_x$  and whose width is  $\Delta t$ . We have shaded in the rectangle representing the 7th term  $v_{x7}\Delta t$ . We see that  $v_{x7}\Delta t$  is just the *area* of the shaded rectangle, and it is clear that the sum of all the areas of the individual rectangles is the total area under the curve, starting at time  $t_i$  and ending at time  $t_f$ . Here we are beginning to see that the process of integration is equivalent to finding the area under a curve.

With a simple curve like the constant velocity  $v_x(t)$  in Figure (11c), we see by inspection that the total area from  $t_i$  to  $t_f$  is just the area of the complete rectangle of height  $v_x$  and width  $(t_f - t_i)$ . Thus

$$\mathbf{s}_{\mathbf{x}} = \mathbf{v}_{\mathbf{x}} \times (\mathbf{t}_{\mathbf{f}} - \mathbf{t}_{\mathbf{i}}) \tag{21}$$

This is the expected result for constant velocity, namely

$$\begin{array}{l} \text{distance} \\ \text{traveled} \end{array} = \ \text{velocity} \times \text{time} \quad \begin{array}{c} \text{for} \\ \text{constant} \\ \text{velocity} \end{array} \quad (21a) \end{array}$$

To see that you are not restricted to the case of constant velocity, suppose you drove on a freeway due east (the x direction) starting at 9:00 AM and stopping for lunch at 12 noon. Every minute during your trip you wrote down the speedometer reading so that you had an accurate plot of  $v_x(t)$  for the entire morning, a plot like that shown in Figure (12). From such a plot, could you determine the distance  $s_x$  that you had travelled?

Your best answer is to multiply each value  $v_i$  of your velocity by the time  $\Delta t$  to calculate the average distance traveled each minute. Summing these up from the initial time  $t_i = 9:00AM$  to the final time  $t_f = noon$ , you have as your estimate

$$s_x \approx \sum_i v_{xi} \Delta t$$

(The symbol  $\approx$  means *approximately equal*.)

To get a more accurate value for the distance traveled, you should measure your velocity at shorter time intervals  $\Delta t$  and add up the larger number of smaller rectangles. The precise answer should be obtained in the limit as  $\Delta t$  goes to zero

$$s_{x} = \lim_{\Delta t \to 0} \sum_{i} v_{xi} \Delta t = \int_{t_{i}}^{t_{f}} v_{x}(t) dt$$
(22)

This limit is just the area under the curve that is supposed to represent the instantaneous velocity  $v_x(t)$ .



Figure 12

Plot of  $v_x(t)$  for a trip starting at 9:00 AM and finishing at noon. The distance traveled is the area under the curve.

Thus we can interpret the integral of a curve as the area under the curve even when the curve is not constant or flat. Mathematicians concern themselves with curves that are so wild that it is difficult or impossible to determine the area under them. Such curves seldom appear in physics problems.

While the basic idea of integration is simple—just finding the area under a curve—in practice it can be quite difficult to calculate the area. Much of an introductory calculus course is devoted to finding the formulas for the areas under various curves. There are also books called *tables of integrals* where you look up the formula for a curve and the table tells you the formula for the area under that curve.

In Chapter 16 of the physics text, we will discuss a mathematical technique called *Fourier analysis*. This is a technique in which we can describe the shape of any continuous curve in terms of a sum of sin waves. (Why we want to do that will become clear then.) The process of Fourier analysis involves finding the area under some very complex curves, curves often involving experimental data for which we have no formula, only graphs. Such curves cannot be integrated by using a table of integrals, with the result that Fourier analysis was not widely used until the advent of the modern digital computer.

The computer made a difference, because we can find the area under almost any curve by breaking the curve into short pieces of length  $\Delta t$ , calculating the area  $v_i \Delta t$ of each narrow rectangle, and adding up the area of the rectangles to get the total area. If the curve is so wild that we have to break it into a million segments to get an accurate answer, that might be too hard to do by hand, but it usually a very simple and rapid job for a computer. Computers can be much more efficient than people at integration.

# **The Process of Integrating**

There is a language for the process of integration which we will now take you through. In each case we will check that the results are what we would expect from our summation definition, or the idea that an integral is the area under a curve.

The simplest integral we will encounter in the calculation of the area under a curve of unit height as shown in Figure (13). We have the area of a rectangle of height 1 and length  $(t_f - t_i)$ 

$$\int_{t_{i}}^{t_{f}} 1 dt = \int_{t_{i}}^{t_{f}} dt = (t_{f} - t_{i})$$
(22)
$$1 = \frac{1}{t_{i}} = 1(t_{f} - t_{i}) = \frac{1}{t_{f}} t$$

**Figure 13** *Area under a curve of unit height.* 

We will use some special language to describe this integration. We will say that the integral of dt is simply the time t, and that the integral of dt from  $t_i$  to  $t_f$  is equal to t evaluated from  $t_i$  to  $t_f$ . In symbols this is written as

$$\int_{t_{i}}^{t_{f}} dt = t \Big|_{t_{i}}^{t_{f}} = (t_{f} - t_{i})$$
(23)

Recall that the vertical line after a variable means to evaluate that variable at the final position  $t_f$  (upper value), minus that variable evaluated at the initial position  $t_i$  (lower value). Notice that this prescription gives the correct answer.

The next simplest integral is the integral of a constant, like a constant velocity  $v_x$  over the interval  $t_i$  to  $t_f$ 



**Figure 14** Area under the constant  $v_r$  curve.

Since  $(t_f - t_i) = \int_{t_i}^{t_f} dt$ , we can replace  $(t_f - t_i)$  in Equation (24) by the integral to get

$$\int_{t_i}^{t_f} v_x dt = v_x \int_{t_i}^{t_f} dt \qquad v_x a \ constant$$
(25)

and we see that a constant like  $v_x$  can be taken outside the integral sign.

Let us try the simplest case we can think of where  $v_x$  is not constant. Suppose  $v_x$  starts at zero at time  $t_i = 0$  and increases linearly according to the formula

$$v_x = at$$
 (26)



When we get up to the time  $t_f$  the velocity will be  $(at_f)$  as shown in Figure (15). The area under the curve  $v_x = at$  is a triangle whose base is of length  $t_f$  and height is  $at_f$ . The area of this triangle is one half the base times the height, thus we get for the distance  $s_x$  traveled by an object moving with this velocity

$$s_{x} = \int_{0}^{t_{f}} v_{x} dt = \frac{1}{2} (base) \times (altitude)$$
  
$$= \frac{1}{2} (t_{f}) (at_{f}) = \frac{1}{2} at_{f}^{2}$$
(27)

Now let us repeat the same calculation using the language one would find in a calculus book. We have

$$s_x = \int_0^{t_f} v_x dt = \int_0^{t_f} (at) dt$$
 (28)

The constant (a) can come outside, and we know that the answer is  $1/2at_f^2$ , thus we can write

$$s_x = a \int_0^{t_f} t dt = \frac{1}{2} a t_f^2$$
 (29)

In Equation (29) we can cancel the a's to get the result

$$\int_{0}^{t_{x}} t dt = \frac{1}{2} t_{f}^{2}$$
(30)

In a calculus text, you would find the statement that the integral  $\int t dt$  is equal to  $t^2/2$  and that the integral should be evaluated as follows

$$\int_{0}^{t_{\rm f}} t dt = \frac{t^2}{2} \Big|_{0}^{t_{\rm f}} = \frac{t_{\rm f}^2}{2} - \frac{0}{2} = \frac{t_{\rm f}^2}{2}$$
(31)

# **Indefinite Integrals**

When we want to measure an actual area under a curve, we have to know where to start and stop. When we put these limits on the integral sign, like  $t_i$  and  $t_f$ , we have what is called a *definite integral*. However there are times where we just want to know what the form of the integral is, with the idea that we will put in the limits later. In this case we have what is called an *indefinite integral*, such as

$$\int t dt = \frac{t^2}{2} \quad indefinite integral \tag{32}$$

The difference between our definite integral in Equation (31) and the indefinite one in Equation (32) is that we have not chosen the limits yet in Equation (32). If possible, a table of integrals will give you a formula for the indefinite integral and let you put in whatever limits you want.

# **Integration Formulas**

For some sets of curves, there are simple formulas for the area under them. One example is the set of curves of the form  $t^n$ . We have already considered the cases where n = 0 and n = 1.



Some results we will prove later are



$$n = 3$$

$$\int t^{3} dt = \frac{t^{4}}{4}$$
(33a,b,c,d)

Looking at the way these integrals are turning out, we suspect that the general rule is

$$\int t^n dt = \frac{t^{n+1}}{n+1}$$
(34)

It turns out that Equation (34) is a general result for any value of n except n = -1. If n = -1, then you would have division by zero, which cannot be the answer. (We will shortly discuss the special case where n = -1.)

As long as we stay away from the n = -1 case, the formula works for negative numbers. For example

$$\int t^{-2} dt = \int \frac{dt}{t^2} = \frac{t^{(-2+1)}}{-2+1} = \frac{t^{-1}}{(-1)}$$

$$\int \frac{dt}{t^2} = -\frac{1}{t}$$
(35)

In our discussion of gravitational and electrical potential energy, we will encounter integrals of the form seen in Equation (35).

#### Exercise 1

Using Equation (34) and the fact that constants can come outside the integral, evaluate the following integrals:



# NEW FUNCTIONS Logarithms

We have seen that when we integrate a curve or function like  $t^2$ , we get a new function  $t^3/3$ . The functions  $t^2$  and  $t^3$  appear to be fairly similar; the integration did not create something radically different. However, the process of integration can lead to some curves with entirely different behavior. This happens, for example, in that special case n = -1 when we try to do the integral of  $t^{-1}$ .

It is certainly not hard to plot  $t^{-1}$ , the result is shown in Figure (16). Also there is nothing fundamentally difficult or peculiar about measuring the area under the  $t^{-1}$  curve from some  $t_i$  to  $t_f$ , as long as we stay away from the origin t = 0 where  $t^{-1}$  blows up. The formula for this area turns out, however, to be the new function called the *natural logarithm*, abbreviated by the symbol *ln*. The area in Figure (16) is given by the formula

$$\int_{t_{i}}^{t_{f}} \frac{1}{t} dt = \ln(t_{f}) - \ln(t_{i})$$
(36)



Plot of  $t^{-1}$ . The area under this curve is the natural logarithm ln.

Two of the important but peculiar features of the natural logarithm are

$$\ln(ab) = \ln(a) + \ln(b) \tag{37}$$

$$\ln(\frac{1}{a}) = -\ln(a) \tag{38}$$

Thus we get, for example

$$\ln(t_{f}) - \ln(t_{i}) = \ln(t_{f}) + \ln\left(\frac{1}{t_{i}}\right)$$
$$= \ln\left(\frac{t_{f}}{t_{i}}\right)$$
(39)

Thus the area under the curve in Figure (16) is

$$\int_{t_i}^{t_f} \frac{dt}{t} = \ln\left(\frac{t_f}{t_i}\right) \tag{40}$$

While the natural logarithm has some rather peculiar properties it is easy to evaluate because it is available on all scientific calculators. For example, if  $t_i = .5$  seconds and  $t_f = 4$  seconds, then we have

$$\ln\left(\frac{t_{\rm f}}{t_{\rm i}}\right) = \ln\left(\frac{4}{.5}\right) = \ln\left(8\right) \tag{41}$$

Entering the number 8 on a scientific calculator and pressing the button labeled *ln*, gives

$$\ln(8) = 2.079 \tag{42}$$

which is the answer.

#### Exercise 2

Evaluate the integrals

$$\int_{.001}^{1000} \frac{dx}{x} \qquad \int_{.000001}^{1} \frac{dx}{x}$$

Why are the answers the same?

#### **The Exponential Function**

We have just seen that, while the logarithm function may have some peculiar properties, it is easy to evaluate using a scientific calculator. The question we now want to consider is whether there is some function that undoes the logarithm. When we enter the number 8 into the calculator and press ln, we get the number 2.079. Now we are asking if, when we enter the number 2.079, can we press some key and get back the number 8? The answer is, you press the key labeled  $e^x$ . The  $e^x$  key performs the *exponential function* which undoes the logarithm function. We say that the exponential function  $e^x$  is the *inverse* of the logarithm function ln.

## **Exponents to the Base 10**

You are already familiar with exponents to the base 10, as in the following examples

$$10^{0} = 1$$
  

$$10^{1} = 10$$
  

$$10^{-1} = 1/10 = .1$$
  

$$10^{2} = 100$$
  

$$10^{-2} = 1/100 = .01$$
  

$$10^{-6} = .000001$$
  
(43)

The exponent, the number written above the 10, tells us how many factors of 10 are involved. A minus sign means how many factors of 10 we divide by. From this alone we deduce the following rules for the exponent to the base 10.

$$10^{-a} = \frac{1}{10^{a}} \tag{44}$$

$$10^{a} \times 10^{b} = 10^{a+b} \tag{45}$$

(Example  $10^2 \times 10^3 = 100 \times 1000 = 100,000$ .)

The inverse of the *exponent to the base 10* is the function called *logarithm to the base 10* which is denoted by the key labeled *log* on a scientific calculator. Formally this means that

$$\log\left(10^{\mathrm{y}}\right) = \mathrm{y} \tag{46}$$

Check this out on your scientific calculator. For example, enter the number 1,000,000 and press the *log* button and see if you get the number 6. Try several examples so that you are confident of the result.

## The Exponential Function y<sup>x</sup>

Another key on your scientific calculator is labeled  $y^x$ . This allows you to determine the value of any number y raised to the power (or exponent) x. For example, enter the number y = 10, and press the  $y^x$  key. Then enter the number x=6 and press the =key. You should see the answer

$$y^{x} = 10^{6} = 1000000$$

It is quite clear that all exponents obey the same rules we saw for powers of 10, namely

$$y^a \times y^b = y^{a+b} \tag{47}$$

(Example  $y^2 \times y^3 = (y \times y)(y \times y \times y) = y^5$ .)

And as before

$$y^{-a} \equiv \frac{1}{y^a} \tag{48}$$

#### **Exercise 3**

Use your scientific calculator to evaluate the following quantities. (You should get the answers shown.)

(a) 10 <sup>6</sup>	(100000)
(b) $2^3$	(8)

(b)  $2^3$  (8) (c)  $23^0$  (1)

(d)  $10^{-1}$  (.1)

(To do this calculation, enter 10, then press  $y^x$ . Then enter 1, then press the +/– key to change it to –1, then press = to get the answer .1)

(e)  $2^{-.5}$  (1/ $\sqrt{2}$ =.707) (f) log (10) (1) (g) ln (2.7183) (1) (very close to 1)

Try some other examples on your own to become completely familiar with the  $y^x$  key. (You should note that any positive number raised to the 0 power is 1. Also, some calculators, in particular the one I am using, cannot handle any negative values of y, not even  $(-2)^2$  which is +4)

#### Euler's Number e = 2.7183...

We have seen that the function *log* on the scientific calculator undoes, is the inverse of, powers of 10. For example, we saw that

 $\log(10^{x}) = x$  (46 repeated)

Example:  $\log(10^6) = 6$ 

Earlier we saw that the exponential function  $e^x$  was the inverse of the natural logarithm ln. This means that

$$\ln\left(e^{x}\right) = x \tag{49}$$

The difference between the logarithm *log* and the natural logarithm *ln*, is that *log* undoes exponents of the number 10, while *ln* undoes exponents of the number *e*. This special number *e*, one of the fundamental mathematical constants like  $\pi$ , is known as *Euler's number*, and is always denoted by the letter *e*.

You can find the numerical value of Euler's number e on your calculator by evaluating

$$e^1 = e \tag{50}$$

To do this, enter 1 into your calculator, press the  $e^x$  key, and you should see the result

$$e^1 = e = 2.718281828$$
 (51)

We will run into this number throughout the course. You should remember that *e* is about 2.7, or you might even remember 2.718. (Only remembering *e* as 2.7 is as klutzy as remembering  $\pi$  as 3.1)

The terminology in math courses is that the function *log*, which undoes exponents of the number 10, is the *logarithm to the base 10*. The function *ln*, what we have called the *natural logarithm*, which undoes exponents of the number *e*, is the *logarithm to the base e*. You can have logarithms to any base you want, but in practice we only use base 10 (because we have 10 fingers) and the base *e*. The base *e* is special, in part because that is the logarithm that naturally arises when we integrate the function 1/x. We will see shortly that the functions *ln* and  $e^x$  have several more, very special features.

### DIFFERENTIATION AND INTEGRATION

The scientific calculator is a good tool for seeing how the functions like ln and  $e^x$  are inverse of each other. Another example of inverse operations is integration and differentiation. We have seen that integration allows us to go the other way from differentiation [finding x(t) from v(t), rather than v(t) from x(t)]. However it is not so obvious that integration and differentiation are inverse operations when you think of integration as finding the area under a curve, and differentiation as finding limits of  $\Delta x/\Delta t$  as  $\Delta t$  goes to zero. It is time now to make this relationship clear.

First, let us review our concept of a derivative. Going back to our strobe photograph of Figure (3), replacing  $\vec{R}_i$  by  $\vec{R}(t)$  and  $\vec{R}_{i+1}$  by  $\vec{R}(t+\Delta t)$ , as shown in Figure (3a), our strobe velocity was then given by

$$\vec{v}(t) = \frac{\vec{R}(t+\Delta t) - \vec{R}(t)}{\Delta t}$$
 (52)

The calculus definition of the velocity is obtained by reducing the strobe time interval  $\Delta t$  until we obtain the instantaneous velocity  $\vec{v}$ .



**Figure 3a** Defining the strobe velocity.

While Equation (53) looks like it is applied to the explicit case of the strobe photograph of projectile motion, it is easily extended to cover any process of differentiation. Whatever function we have [we had  $\vec{R}(t)$ , suppose it is now f(t)], evaluate it at two closely spaced times, subtract the older value from the newer one, and divide by the time difference  $\Delta t$ . Taking the limit as  $\Delta t$  becomes very small gives us the derivative

$$\frac{\mathrm{d}f(t)}{\mathrm{d}t} \equiv \lim_{\Delta t \to 0} \frac{f(t + \Delta t) - f(t)}{\Delta t}$$
(54)

The variable with which we are differentiating does not have to be time t. It can be any variable that we can divide into small segments, such as x;

$$\frac{d}{dx}f(x) \equiv \lim_{\Delta x \to 0} \frac{f(x + \Delta x) - f(x)}{\Delta x}$$
(55)

Let us see how the operation defined in Equation (55) is the inverse of finding the area under a curve.

Suppose we have a curve, like our old  $v_x(t)$  graphed as a function of time, as shown in Figure (17). To find out how far we traveled in a time interval from  $t_i$  to some later time T, we would do the integral

$$\mathbf{x}(\mathbf{T}) = \int_{t_i}^{\mathbf{T}} \mathbf{v}_{\mathbf{x}}(t) \, \mathrm{d}t \tag{56}$$

The integral in Equation (56) tells us how far we have gone at any time T during the trip. The quantity x(T) is a function of this time T.



**Figure 17** *The distance traveled by the time T is the area under the velocity curve up to the time T.* 

Now let us differentiate the function x(T) with respect to the variable T. By our definition of differentiation we have

$$\frac{d}{dT}x(T) = \lim_{\Delta T \to 0} \frac{x(T + \Delta T) - x(T)}{\Delta T}$$
(57)

Figure (17) shows us the function x(T). It is the area under the curve v(t) starting at  $t_i$  and going up to time t=T. Figure (18) shows us the function  $x(T + \Delta T)$ . It is the area under the same curve, starting at  $t_i$  but going up to  $t = T + \Delta t$ . When we subtract these two areas, all we have left is the area of the slender rectangle shown in Figure (19).



**Figure 17 repeated** *The distance* x(T) *traveled by the time* T



Figure 18

The distance  $x(T+\Delta t)$  traveled by the time  $T+\Delta t$ .



The distance  $x(T+\Delta t) - x(T)$ traveled during the time  $\Delta t$ .

The rectangle has a height approximately v(T) and a width  $\Delta T$  for an area

$$x(T + \Delta T) - x(T) = v_x(T)\Delta t$$
(58)

Dividing through by  $\Delta T$  gives

$$v_{x}(T) = \frac{x(T + \Delta T) - x(T)}{\Delta T}$$
(59)

The only approximation in Equation (59) is at the top of the rectangle. If the curve is not flat,  $v_x(T + \Delta T)$  will be different from  $v_x(T)$  and the area of the sliver will have a value somewhere between  $v_x(T)\Delta t$  and  $v_x(T + \Delta t)\Delta t$ . But if we take the limit as  $\Delta T$  goes to zero, the value of  $v_x(T + \Delta T)$  must approach  $v_x(T)$ , and we end up with the exact result

$$v_{x}(T) = \lim_{\Delta t \to 0} \frac{x(T + \Delta t) - x(T)}{\Delta t}$$
(60)

This is just the derivative dx(t)/dt evaluated at t = T.

$$\left| \mathbf{v}_{\mathbf{x}}(\mathbf{T}) = \frac{\mathbf{d}\mathbf{x}(\mathbf{t})}{\mathbf{d}\mathbf{t}} \right|_{\mathbf{t} = \mathbf{T}}$$
(61a)

where we started from

$$\mathbf{x}(\mathbf{T}) = \int_{t_i}^{\mathbf{T}} \mathbf{v}_{\mathbf{x}}(t) \, \mathrm{d}t$$
 (61b)

Equations (61a) and (61b) demonstrate explicitly how differentiation and integration are inverse operations. The derivative allowed us to go from x(t) to  $v_x(t)$  while the integral took us from  $v_x(t)$  to x(t). This inverse is not as simple as pushing a button on a calculator to go from *ln* to  $e^x$ . Here we have to deal with limits on the integration and a shift of variables from t to T. But these two processes do allow us to go back and forth.

#### A Fast Way to go Back and Forth

We introduced our discussion of integration by pointing out that equations

$$v_x(t) = \frac{dx(t)}{dt};$$
  $a_x(t) = \frac{dv_x(t)}{dt}$  (62a,b)

went the wrong way in that we were more likely to know the acceleration  $a_x(t)$  and from that want to calculate the velocity  $v_x(t)$  and distance traveled x(t). After many steps, we found that integration was what we needed.

We do not want to repeat all those steps. Instead we would like a quick and simple way to go the other way around. Here is how you do it. Think of the dt in (62a) as a small but finite time interval. That means we can treat it like any other number and multiply both sides of Equation (62a) through by it.

$$v_{x}(t) = \frac{dx(t)}{dt}$$
$$dx(t) = v_{x}(t)dt$$
(63)

Now integrate both sides of Equation (63) from some initial time  $t_i$  to a final time T. (If you do the same thing to both sides of an equation, both sides should still be equal to each other.)

$$\int_{t_i}^{T} dx(t) = \int_{t_i}^{T} v_x(t) dt$$
(64)

If dt is to be thought of as a small but finite time step, then dx(t) is the small but finite distance we moved in the time dt. The integral on the left side of Equation (64) is just the sum of all these short distances moved, which is just the total distance moved during the time from  $t_i$  to T.

$$\int_{t_{i}}^{T} dx(t) = x(t) \Big|_{t_{i}}^{T} = x(T) - x(t_{i})$$
(65)

Thus we end up with the result

$$\mathbf{x}(t)\Big|_{t_{i}}^{\mathrm{T}} = \int_{t_{i}}^{\mathrm{T}} \mathbf{v}_{\mathbf{x}}(t)dt$$
(66)

Equation (66) is a little more general than (62b) for it allows for the fact that  $x(t_i)$  might not be zero. If,

however, we say that we started our trip at  $x(t_i) = 0$ , then we get the result

$$x(T) = \int_{t_1}^{T} v_x(t) dt$$
(67)

representing the distance traveled since the start of the trip.

# **Constant Acceleration Formulas**

The constant acceleration formulas, so well known from high school physics courses, are an excellent application of the procedures we have just described.

We will begin with motion in one dimension. Suppose a car is traveling due east, in the x direction, and for a while has a constant acceleration  $a_x$ . The car passes us at a time  $t_i = 0$ , traveling at a speed  $v_{x0}$ . At some later time T, if the acceleration  $a_x$  remains constant, how far away from us will the car be?

We start with the equation

$$a_{x}(t) = \frac{dv_{x}(t)}{dt}$$
(68)

Multiplying through by dt to get

 $dv_x(t) = a_x(t)dt$ 

then integrating from time  $t_i = 0$  to time  $t_f = T$ , we get

$$\int_0^T dv_x(t) = \int_0^T a_x(t)dt$$
(69)

Since the integral  $\int dv_x(t) = v_x(t)$ , we have

$$\int_{0}^{T} dv_{x}(t) = v_{x}(t) \Big|_{0}^{T} = v_{x}(T) - v_{x}(0)$$
(70)

where  $v_x(0)$  is the velocity  $v_{x0}$  of the car when it passed us at time t = 0.

While we can always do the left hand integral in Equation (69), we cannot do the right hand integral until we know  $a_x(t)$ . For the constant acceleration problem, however, we know that  $a_x(t) = a_x$  is constant, and we have

$$\int_0^T a_x(t)dt = \int_0^T a_x dt$$
(71)

Since constants can come outside the integral sign, we get

$$\int_{0}^{T} a_{x} dt = a_{x} \int_{0}^{T} dt = a_{x} t \Big|_{0}^{T} = a_{x} T$$
 (72)

where we used  $\int dt = t$ . Substituting Equations (70) and (72) in (69) gives

$$\mathbf{v}_{\mathbf{x}}\mathbf{T} - \mathbf{v}_{\mathbf{x}\mathbf{0}} = \mathbf{a}_{\mathbf{x}}\mathbf{T} \tag{73}$$

Since Equation (73) applies for any time T, we can replace T by t to get the well known result

$$\mathbf{v}_{\mathbf{x}}(t) = \mathbf{v}_{\mathbf{x}0} + \mathbf{a}_{\mathbf{x}}t \qquad (\mathbf{a}_{\mathbf{x}} \textit{ constant}) \tag{74}$$

Equation (74) tells us the speed of the car at any time t after it passed us, as long as the acceleration remains constant.

To find out how far away the car is, we start with the equation

$$v_{x}(t) = \frac{dx(t)}{dt}$$
(62a)

Multiplying through by dt to get

 $dx(t) = v_x(t) dt$ 

then integrating from time t = 0 to time t = T gives (as we saw earlier)

$$\int_{0}^{T} dx(t) = \int_{0}^{T} v_{x}(t) dt$$
 (75)

The left hand side is

$$\int_{0}^{T} dx(t) = x(t) \Big|_{0}^{T} = x(T) - x(0)$$
(76)

If we measure along the x axis, starting from where we are (where the car was at t = 0) then x(0) = 0.

In order to do the right hand integral in Equation (75), we have to know what the function  $v_x(t)$  is. But for constant acceleration, we have from Equation (74)  $v_x(t) = v_{x0} + a_x t$ , thus

$$\int_{0}^{T} v_{x}(t) dt = \int_{0}^{T} (v_{x0} + a_{x}t) dt$$
 (77)

One of the results of integration that you should prove for yourself (just sketch the areas) is the rule

$$\int_{i}^{f} \left[ a(x) + b(x) \right] dx = \int_{i}^{f} a(x) dx + \int_{i}^{f} b(x) dx \quad (78)$$

thus we get

$$\int_{0}^{T} (v_{x0} + a_x t) dt = \int_{0}^{T} v_{x0} dt + \int_{0}^{T} a_x t dt \quad (79)$$

Since constants can come outside the integrals, this is equal to

$$\int_{0}^{T} (v_{x0} + a_{x}t)dt = v_{x0} \int_{0}^{T} dt + a_{x} \int_{0}^{T} t dt \quad (80)$$

Earlier we saw that

$$\int_{0}^{T} dt = t \Big|_{0}^{T} = T - 0 = T$$
 (23)

$$\int_{0}^{T} t dt = \frac{t^{2}}{2} \bigg|_{0}^{T} = \frac{T^{2}}{2} - 0 = \frac{T^{2}}{2}$$
(30)

Thus we get

$$\int_{0}^{T} (v_{x0} + a_x t) dt = v_{x0} T + \frac{1}{2} a_x T^2$$
(81)

Using Equations (76) and (81) in (75) gives

$$x(T) - x_0 = v_{x0}T + \frac{1}{2}a_xT^2$$

Taking  $x_0 = 0$  and replacing T by t gives the other constant acceleration formula

$$\mathbf{x}(t) = \mathbf{v}_{\mathbf{x}0}t + \frac{1}{2}\mathbf{a}_{\mathbf{x}}t^2 \qquad (\mathbf{a}_{\mathbf{x}} \textit{ constant}) \qquad (82)$$

You can now see that the factor of  $t^2/2$  in the constant acceleration formulas comes from the integral  $\int t dt$ .

#### **Exercise 4**

Find the formula for the velocity v(t) and position x(t) for a car moving with constant acceleration  $a_x$ , that was located at position  $x_i$  at some initial time  $t_i$ .

Start your calculation from the equations

$$v_{x}(t) = \frac{dx(t)}{dt}$$
$$a_{x}(t) = \frac{dv_{x}(t)}{dt}$$

and go through all the steps that we did to get Equations (74) and (82). See if you can do this without looking at the text.

If you have to look back to see what some steps are, then finish the derivation looking at the text. Then a day or so later, clean off your desk, get out a blank sheet of paper, write down this problem, put the book away and do the derivation. Keep doing this until you can do the derivation of the constant acceleration formulas without looking at the text.

## **Constant Acceleration Formulas** in Three Dimensions

To handle the case of motion with constant acceleration in three dimensions, you start with the separate equations

$$v_{x}(t) = \frac{dx(t)}{dt} \qquad a_{x}(t) = \frac{dv_{x}(t)}{dt}$$
$$v_{y}(t) = \frac{dy(t)}{dt} \qquad a_{y}(t) = \frac{dv_{y}(t)}{dt}$$
$$v_{z}(t) = \frac{dz(t)}{dt} \qquad a_{z}(t) = \frac{dv_{z}(t)}{dt} \qquad (83)$$

Then repeat, for each pair of equations, the steps that led to the constant acceleration formulas for motion in the x direction. The results will be

$$\begin{aligned} x(t) &= v_{x0}t + \frac{1}{2}a_{x}t^{2} & v_{x}(t) = v_{x0} + a_{x}t \\ y(t) &= v_{y0}t + \frac{1}{2}a_{y}t^{2} & v_{y}(t) = v_{y0} + a_{y}t \quad (84) \\ z(t) &= v_{z0}t + \frac{1}{2}a_{z}t^{2} & v_{z}(t) = v_{z0} + a_{z}t \end{aligned}$$

The final step is to combine these six equations into the two vector equations

$$\vec{x}(t) = \vec{v}_0 t + \frac{1}{2} \vec{a} t^2$$
;  $\vec{v}(t) = \vec{v}_0 + \vec{a} t$  (85)

These are the equations we analyzed graphically in Chapter 3 of the physics text, in Figure (3-34) and Exercise (3-9). (There we wrote  $\vec{s}$  instead of  $\vec{x}(t)$ , and  $\vec{v}_i$ , rather than  $v_0$ .)

In many introductory physics courses, considerable emphasis is placed on solving constant acceleration problems. You can spend weeks practicing on solving these problems, and become very good at it. However, when you have done this, you have not learned very much physics because most forms of motion are not with constant acceleration, and thus the formulas do not apply. The formulas were important historically, for they were the first to allow the accurate prediction of motion (of cannonballs). But if too much emphasis is placed on these problems, students tend to use them where they do not apply. For this reason we have placed the exercises using the constant acceleration equations in an appendix at the end of chapter 4 of the physics text. There are plenty of problems there for all the practice you will need with these equations. Doing these exercises requires only algebra, there is no practice with calculus. To get some experience with calculus, be sure that you can confidently do Exercise 4.

## MORE ON DIFFERENTIATION

In our discussion of integration, we saw that the basic idea was that the integral of some curve or function f(t) was equal to the area under that curve. That is an easy enough concept. The problems arose when we actually tried to find the formulas for the areas under various curves. The only areas we actually calculated were the rectangular area under f(t) =constant and the triangular area under f(t) =at. It was perhaps a surprise that the area under the simple curve 1/t should turn out to be a logarithm.

For differentiation, the basic idea of the process is given by the formula

$$\frac{\mathrm{d}f(t)}{\mathrm{d}t} = \lim_{\Delta t \to 0} \frac{f(t + \Delta t) - f(t)}{\Delta t} \qquad (54 \text{ repeated})$$

Equation (54) is short hand notation for a whole series of steps which we introduced through the use of strobe photographs. The basic idea of differentiation is more complex than integration, but, as we will now see, it is often a lot easier to find the derivative of a curve than its integral.

## **Series Expansions**

An easy way to find the formula for the derivative of a curve is to use a series expansion. We will illustrate the process by using the binomial expansion to calculate the derivative of the function  $x^n$  where n is any constant.

We used the binomial expansion, or at least the first two terms, in Chapter 1 of the physics text. That was during our discussion of the approximation formulas that are useful in relativistic calculations. As we mentioned in Exercise (1-5), the binomial expansion is

$$(x + \alpha)^{n} = x^{n} + n\alpha x^{n-1} + \frac{n(n-1)}{2!} \alpha^{2} x^{n-2} \cdots$$
(86)

When  $\alpha$  is a number much smaller than 1 ( $\alpha < < 1$ ), we can neglect  $\alpha^2$  compared to  $\alpha$  (if  $\alpha = .01$ ,  $\alpha^2 = .0001$ ), with the result that we can accurately approximate  $(x + \alpha)^n$  by

$$(x + \alpha)^n \approx x^n + n\alpha x^{n-1} \qquad \alpha << 1 \qquad (87)$$

Equation (87) gives us all the approximation formulas found in Equations (1-20) through (1-25) on page 1-28 of the physics text.

As an example of Equation (87), just to see that it works, let us take x = 5, n = 7 and  $\alpha = .01$  to calculate  $(5.01)^7$ . From the calculator we get

$$(5.01)^7 = 79225.3344 \tag{88}$$

(To do this enter 5.01, press the  $y^x$  button, then enter 7 and press the = button.) Let us now see how this result compares with

$$(x + \alpha)^{n} \approx x^{n} + n\alpha x^{n-1}$$

$$(5 + .01)^{7} \approx 5^{7} + 7(.01)5^{6}$$
(89)

We have

$$5^7 = 78125$$
 (90)

 $7 \times .01 \times 5^{6} = 7 \times .01 \times 15625 = 1093.75$  (91)

Adding the numbers in (90) and (91) together gives

$$5^7 + 7(.01)5^6 = 79218.75 \tag{92}$$

Thus we end up with 79218 instead of 79225, which is not too bad a result. The smaller  $\alpha$  is compared to one, the better the approximation.

# Derivative of the Function x<sup>n</sup>

We are now ready to use our approximation formula (87) to calculate the derivative of the function  $x^n$ . From the definition of the derivative we have

$$\frac{d(x^{n})}{dx} = \lim_{\Delta x \to 0} \frac{(x + \Delta x)^{n} - x^{n}}{\Delta x}$$
(93)

Since  $\Delta x$  is to become infinitesimally small, we can use our approximation formula for  $(x + \alpha)^n$ . We get

$$(x + \alpha)^{n} \approx x^{n} + n(\alpha)x^{n-1} \qquad (\alpha << 1)$$
$$(x + \Delta x)^{n} \approx x^{n} + n(\Delta x)x^{n-1} \qquad (\Delta x << 1) (94)$$

Using this in Equation (93) gives

$$\frac{d(x^{n})}{dx} = \lim_{\Delta x \to 0} \left[ \frac{\left( x^{n} + n(\Delta x)x^{n-1} \right) - x^{n}}{\Delta x} \right] \quad (95)$$

We used an equal sign rather than an approximately equal sign in Equation (95) because our approximation formula (94) becomes exact when  $\Delta x$  becomes infinitesimally small.

In Equation (95), the terms  $x^n$  cancel and we are left with

$$\frac{d(x^{n})}{dx} = \lim_{\Delta x \to 0} \left[ \frac{n(\Delta x)x^{n-1}}{\Delta x} \right]$$
(96)

At this point, the factors  $\Delta x$  cancel and we have

$$\frac{d(x^{n})}{dx} = \lim_{\Delta x \to 0} \left[ nx^{n-1} \right]$$
(97)

Since no  $\Delta x$ 's remain in our formula, we end up with the exact result

$$\frac{\mathrm{d}(\mathbf{x}^{\mathrm{n}})}{\mathrm{d}\mathbf{x}} = \mathrm{n}\mathbf{x}^{\mathrm{n}-1} \tag{98}$$

Equation (98) is the general formula for the derivative of the function  $x^n$ .

In our discussion of integration, we saw that a constant could come outside the integral. The same thing happens with a derivative. Consider, for example,

$$\frac{d}{dx} \Big[ a f(x) \Big] = \lim_{\Delta x \to 0} \left| \frac{a f(x + \Delta x) - a f(x)}{\Delta x} \right|$$

Since the constant a has nothing to do with the limiting process, this can be written

$$\frac{d}{dx} \left[ af(x) \right] = a \lim_{\Delta x \to 0} \left[ \frac{f(x + \Delta x) - f(x)}{\Delta x} \right]$$
$$= a \frac{df(x)}{dx}$$
(99)

#### **Exercise 5**

Calculate the derivative with respect to x (i.e., d/dx) of the following functions. (When negative powers of x are involved, assume x is not equal to zero.)

- (a) x
- (b) x<sup>2</sup>
- (C) X<sup>3</sup>
- (d)  $5x^2 3x$

(Before you do part (d), use the definition of the derivative to prove that  $\frac{d}{dx}[f(x) + g(x)] = \frac{df(x)}{dx} + \frac{dg(x)}{dx}$ )

- (e) x<sup>-1</sup>
- (f) x<sup>-2</sup>
- (g) √x
- (h) 1/√x
- (i) 3x<sup>.73</sup>
- (j) 7x<sup>-.2</sup>
- (k) 1

(In part (k) first show that this should be zero from the definition of the derivative. Then write  $1 = x^0$  and show that Equation (98) also works, as long as x is not zero.)

(I) 5

# The Chain Rule

There is a simple trick called the *chain rule* that makes it easy to differentiate a wide variety of functions. The rule is

$$\frac{df[y(x)]}{dx} = \frac{df(y)}{dy}\frac{dy}{dx} \qquad chain rule \quad (100)$$

To see how this rule works, consider the function

$$\mathbf{f}(\mathbf{x}) = \left(\mathbf{x}^2\right)^n \tag{101}$$

We know that this is just  $f(x) = x^{2n}$ , and the derivative is

$$\frac{df(x)}{dx} = \frac{d}{dx} (x^{2n}) = 2nx^{2n-1}$$
(102)

But suppose that we did not know this trick, and therefore did not know how to differentiate  $(x^2)^n$ . We do, however, know how to differentiate powers like  $x^2$  and  $y^n$ . The chain rule allows us to use this knowledge in order to figure out how to differentiate the more complex function $(x^2)^n$ .

We begin by defining y(x) as

$$\mathbf{y}(\mathbf{x}) = \mathbf{x}^2 \tag{103}$$

Then our function  $f(x) = (x^2)^n$  can be written in terms of y as follows

$$f(x) = (x^{2})^{n} = (y(x))^{n} = (y)^{n} = f(y)$$
  
$$f(y) = (y)^{n}$$
 (104)

Differentiating (103) and (104) gives

$$\frac{\mathrm{d}y(x)}{\mathrm{d}x} = \frac{\mathrm{d}}{\mathrm{d}x} \left(x^2\right) = 2x \tag{105}$$

$$\frac{\mathrm{d}f(y)}{\mathrm{d}y} = \frac{\mathrm{d}}{\mathrm{d}y}(y^n) = ny^{n-1} \tag{106}$$

Using (104) and (105) in the chain rule (100) gives

$$\frac{df(y)}{dx} = \frac{df}{dy} \times \frac{dy}{dx} = (ny^{n-1}) \times (2x)$$

$$= 2ny^{n-1}x$$

$$= 2n(x^2)^{n-1}x$$

$$= 2n(x^{2(n-1)})x$$

$$= 2n(x^{(2n-2)})x \qquad (107)$$

$$= 2n(x^{(2n-2)+1})$$

$$= 2nx^{2n-1}$$

which is the answer we expect.

In our example, using the chain rule was more difficult than differentiating directly because we already knew how to differentiate  $x^{2n}$ . But we will shortly encounter examples of new functions that we do not know how to differentiate directly, but which can be written in the form f[y(x)]; and where we know df/dy and dy/dx. We can then use the chain rule to evaluate the derivative df/dx. We will give you practice with the chain rule when we encounter these functions.

# **Remembering The Chain Rule**

The chain rule can be remembered by thinking of the dy's as cancelling as shown.

$$\frac{df(y)}{d\chi}\frac{d\chi}{dx} = \frac{df(y)}{dx}$$
 remembering  
the chain rule (108)

#### Partial Proof of the Chain Rule (optional)

The proof of the chain rule is closely related to cancellation we showed in Equation (108). A partial proof of the rule proceeds as follows.

Suppose we have some function f(y) where y is a function of the variable x. As a result f[y(x)] is itself a function of x and can be differentiated with respect to x.

$$\frac{\mathrm{d}}{\mathrm{d}x} f[y(x)] = \lim_{\Delta x \to 0} \frac{f[y(x + \Delta x)] - f[y(x)]}{\Delta x} \quad (123)$$

Now define the quantity  $\Delta y$  by

$$\Delta y \equiv y(x + \Delta x) - y(x) \tag{124}$$

so that

$$y(x + \Delta x) = y(x) + \Delta y$$
$$f[y(x + \Delta x)] = f(y + \Delta y)$$

and Equation (123) becomes

$$\frac{d}{dx}f[y(x)] = \lim_{\Delta x \to 0} \frac{f(y + \Delta y) - f(y)}{\Delta x}$$
(125)

Now multiply (125) through by

$$1 = \frac{\Delta y}{\Delta y} = \frac{y(x + \Delta x) - y(x)}{\Delta y}$$
(126)

to get

$$\frac{d}{dx} f[y(x)]$$

$$= \lim_{\Delta x \to 0} \left[ \frac{f(y + \Delta y) - f(y)}{\Delta x} \times \frac{y(x + \Delta x) - y(x)}{\Delta y} \right]$$

$$= \lim_{\Delta x \to 0} \left[ \frac{f(y + \Delta y) - f(y)}{\Delta y} \times \frac{y(x + \Delta x) - y(x)}{\Delta x} \right]$$

(127)

where we interchanged  $\Delta x$  and  $\Delta y$  in the denominator.

(We call this a partial proof for the following reason. For some functions y(x), the quantity  $\Delta y = y(x + \Delta x) - y(x)$  may be identically zero for a small range of  $\Delta x$ . In that case we would be dividing by zero (the  $1/\Delta y$ ) even before we took the limit as  $\Delta x$  goes to zero. A more complete proof handles the special cases separately. The resulting chain rule still works however, even for these special cases.)

Since  $\Delta y = y(x + \Delta x) - y(x)$  goes to zero as  $\Delta x$  goes to zero, we can write Equation (127) as

$$\frac{d}{dx} f[y(x)]$$

$$= \begin{bmatrix} \liminf_{\Delta y \to 0} \frac{f(y + \Delta y) - f(y)}{\Delta y} \end{bmatrix}$$

$$\times \begin{bmatrix} \liminf_{\Delta x \to 0} \frac{y(x + \Delta x) - y(x)}{\Delta x} \end{bmatrix}$$

$$= \begin{bmatrix} \frac{df(y)}{dy} \end{bmatrix} \begin{bmatrix} \frac{dy}{dx} \end{bmatrix} \quad (100 \text{ prand})$$

This rule works as long as the derivatives df/dy and dy/dx are meaningful, i.e., we stay away from kinks or discontinuities in f and y.

# **INTEGRATION FORMULAS**

Knowing the formula for the derivative of the function  $x^n$ , and knowing that integration undoes differentiation, we can now use Equation (98)

$$\frac{\mathrm{d}x^n}{\mathrm{d}x} = nx^{n-1}$$

(98epeated)

to find the integral of the function  $x^n$ . We will see that this trick works for all cases except the special case where n = -1, i.e., the special case where the integral is a natural logarithm.

To integrate  $x^n$ , let us go back to our calculation of the distance  $s_x$  or x(t) traveled by an object moving in the x direction at a velocity  $v_x$ . This was given by Equations (19) or (56) as

$$x(t)\Big|_{t_{i}}^{T} = \int_{t_{i}}^{T} v_{x}(t) dt$$
 (128)

where the instantaneous velocity  $v_x(t)$  is defined as

$$v_{x}(t) = \frac{dx(t)}{dt}$$
(129)

Suppose x(t) had the special form

$$\mathbf{x}(t) = t^{n+1} \qquad (a \ special \ case) \tag{130}$$

then we know from our derivative formulas that

$$v(t) = \frac{dx(t)}{dt} = \frac{dt^{(n+1)}}{dt} = (n+1)t^n$$
 (131)

Substituting  $x(t) = t^{n+1}$  and  $v(t) = (n+1)t^n$  into Equation (128) gives

$$x(t)\Big|_{t_{i}}^{T} = \int_{t_{i}}^{T} v_{x}(t) dt$$
 (128)

$$t^{n+1}\Big|_{t_{i}}^{T} = \int_{t_{i}}^{T} (n+1)t^{n}dt$$
$$= (n+1)\int_{t_{i}}^{T} t^{n}dt$$
(132)

Dividing through by (n+1) gives

$$\int_{t_{i}}^{T} t^{n} dt = \frac{1}{n+1} t^{n+1} \Big|_{t_{i}}^{T}$$
(133)

If we choose  $t_i = 0$ , we get the simpler result

$$\int_{0}^{T} t^{n} dt = \frac{T^{n+1}}{n+1}$$
(134)

and the indefinite integral can be written

$$\int t^{n} dt = \frac{t^{n+1}}{n+1}$$
(135)(also34)

This is the general rule we stated without proof back in Equation (34). Note that this formula says nothing about the case n = -1, i.e., when we integrate  $t^{-1} = 1/t$ , because n+1=-1+1=0 and we end up with division by zero. But for all other values of n, we now have derived a general formula for finding the area under any curve of the form  $x^n$  (or  $t^n$ ). This is a rather powerful result considering the problems one encounters actually finding areas under curves. (If you did not do Exercise 1, the integration exercises on page 14, or had difficulty with them, go back and do them now.)

#### **Derivative of the Exponential Function**

The previous work shows us that if we have a series expansion for a function, it is easy to obtain a formula for the derivative of the function. We will now apply this technique to calculate the derivative and integral of the exponential function  $e^x$ .

There is a series expansion for the function  $e^x$  that works for any value of  $\alpha$  in the range -1 to +1.

$$e^{\alpha} \approx 1 + \alpha + \frac{\alpha^2}{2!} + \frac{\alpha^3}{3!} + \cdots$$
 (136)

where  $2! = 2 \times 1$ ,  $3! = 3 \times 2 \times 1 = 6$ , etc. (The quantities 2!, 3! are called *factorials*. For example 3! is called *three factorial*.)

To see how well the series (136) works, consider the case  $\alpha = .01$ . From the series we have, up to the  $\alpha^3$  term

$$\alpha = .01$$
  
 $\alpha^2 = .0001$ ;  $\alpha^2/2 = .00005$   
 $\alpha^3 = .000001$ ;  $\alpha^3/6 = .000000167$ 

Giving us the approximate value

$$1 + \alpha + \frac{\alpha^2}{2!} + \frac{\alpha^3}{3!} = 1.010050167$$
 (137)

When we enter .01 into a scientific calculator and press the  $e^x$  button, we get exactly the same result. Thus the calculator is no more accurate than including the  $\alpha^3$ term in the series, for values of  $\alpha$  equal to .01 or less. Let us now see how to use the series 136 for calculating the derivative of  $e^x$ . We have, from the definition of a derivative,

$$\frac{d}{dx}f(x) \equiv \lim_{\Delta x \to 0} \frac{f(x + \Delta x) - f(x)}{\Delta x}$$
 (56epeat)

If  $f(x) = e^x$ , we get

$$\frac{d(e^{x})}{dx} = \lim_{\Delta x \to 0} \left[ \frac{e^{x + \Delta x} - e^{x}}{\Delta x} \right]$$
(138)

To do this calculation, we have to evaluate the quantity  $e^{x + \Delta x}$ . First, we use the fact that for exponentials

 $e^{a+b} = e^a e^b$ 

(Remember that  $10^{2+3} = 10^2 \times 10^3 = 10^5$ .) Thus

$$e^{x + \Delta x} = e^{x} e^{\Delta x} \tag{139}$$

Now use the approximation formula (136), setting  $\alpha = \Delta x$  and throwing out the  $\alpha^2$  and  $\alpha^3$  and higher terms because we are going to let  $\Delta x$  go to zero

$$e^{\Delta x} \approx 1 + \Delta x \tag{140}$$

Substituting (140) in (139) gives

$$e^{x+\Delta x} \approx e^{x}(1+\Delta x)$$
  
=  $e^{x} + e^{x}\Delta x$  (141)

Next use (141) in (138) to get

$$\frac{d(e^{x})}{dx} = \lim_{\Delta x \to 0} \left[ \frac{(e^{x} + e^{x} \Delta x) - e^{x}}{\Delta x} \right]$$
(142)

The  $e^x$  terms cancel and we are left with

$$\frac{d(e^{x})}{dx} = \lim_{\Delta x \to 0} \left[ \frac{e^{x} \Delta x}{\Delta x} \right] = \lim_{\Delta x \to 0} e^{x}$$
(143)

Since the  $\Delta x's$  cancelled, we are left with the exact result

$$\frac{d(e^{x})}{dx} = e^{x}$$
(144)

We see that the exponential function  $e^x$  has the special property that it is its own derivative.

We will often want to know the derivative, not just of the function  $e^x$  but of the slightly more general result  $e^{ax}$  where a is a constant. That is, we want to find

$$\frac{\mathrm{d}}{\mathrm{dx}}\mathrm{e}^{\mathrm{ax}} \qquad (a = constant) \tag{145}$$

Solving this problem provides us with our first meaningful application of the chain rule

$$\frac{df(y)}{dx} = \frac{df(y)}{dy}\frac{dy}{dx}$$
 (100epeated)

If we set

y≒ax

then we have

$$\frac{\mathrm{d}e^{\mathrm{a}x}}{\mathrm{d}x} = \frac{\mathrm{d}e^{\mathrm{y}}}{\mathrm{d}\mathrm{y}}\frac{\mathrm{d}\mathrm{y}}{\mathrm{d}\mathrm{x}} \tag{147}$$

(146)

Now

$$\frac{\mathrm{d}\mathrm{e}^{\mathrm{y}}}{\mathrm{d}\mathrm{y}} = \mathrm{e}^{\mathrm{y}} \tag{148}$$

$$\frac{dy}{dx} = \frac{d}{dx}(ax) = a\frac{dx}{dx} = a \times 1 = a \qquad (149)$$

Using (148) and (149) in (147) gives

$$\frac{\mathrm{d}\mathrm{e}^{\mathrm{ax}}}{\mathrm{d}\mathrm{x}} = (\mathrm{e}^{\mathrm{y}})(\mathrm{a}) = (\mathrm{e}^{\mathrm{ax}})(\mathrm{a}) = \mathrm{a}\mathrm{e}^{\mathrm{ax}}$$

Thus we have

$$\frac{\mathrm{d}}{\mathrm{d}x}\mathrm{e}^{\mathrm{a}x} = \mathrm{a}\mathrm{e}^{\mathrm{a}x} \tag{150}$$

This result will be used so often it is worth memorizing.

#### **Exercise 6**

For further practice with the chain rule, show that

$$\frac{de^{ax^2}}{dx} = 2axe^{ax^2}$$

Do this by choosing  $y = ax^2$ , and then do it again by choosing  $y = x^2$ .

# Integral of the Exponential Function

To calculate the integral of  $e^{ax}$ , we will use the same trick as we used for the integral of  $x^n$ , but we will be a bit more formal this time. Let us start with Equation (128) relating position x(t) and velocity v(t) = dx(t)/dt go get

$$x(t)\Big|_{t_{i}}^{t_{f}} = \int_{t_{i}}^{t_{f}} v_{x}(t) dt = \int_{t_{i}}^{t_{f}} \frac{dx(t)}{dt} dt \qquad (128)$$

Since Equation (128) holds for any function x(t) [we did not put any restrictions on x(t)], we can write Equation (128) in a more abstract way relating any function f(x) to its derivative df(x)/dx;

$$f(x)\Big|_{x_{i}}^{x_{f}} = \int_{x_{i}}^{x_{f}} \frac{df(x)}{dx} dx$$
(151)

To calculate the integral of  $e^{ax}$ , we set  $f(x) = e^{ax}$  and  $df(x)/dx = ae^{ax}$  to get

$$\left. e^{ax} \right|_{x_i}^{x_f} = \int_{x_i}^{x_f} a e^{ax} dx \tag{152}$$

Dividing (157) through by (a) gives us the definite integral

$$\int_{x_{i}}^{x_{f}} e^{ax} dx = \frac{1}{a} e^{ax} \Big|_{x_{i}}^{x_{f}} \qquad (a = constant) \quad (153)$$

The corresponding indefinite integral is

$$\int e^{ax} dx = \frac{e^{ax}}{a} \qquad (a = constant) \qquad (154)$$

#### Exercise 7

The natural logarithm is defined by the equation

Use Equation (151) to show that

 $\ln(x) = \int \left(\frac{1}{x}\right) dx$ 

$$\frac{\mathrm{d}}{\mathrm{d}x}(\ln x) = \frac{1}{x} \tag{155}$$

(Hint—integrate both sides of Equation (155) with respect to x.)

#### **DERIVATIVE AS THE SLOPE OF A CURVE**

Up to now, we have emphasized the idea that the derivative of a function f(x) is given by the limiting process

$$\frac{df(x)}{dx} = \lim_{\Delta x \to 0} \left| \frac{f(x + \Delta x) - f(x)}{\Delta x} \right| (55 \text{ repeated})$$

We saw that this form was convenient when we had an explicit way of calculating  $f(x + \Delta x)$ , as we did by using a series expansion. However, a lot of words are required to explain the steps involved in doing the limiting process indicated in Equation (55). In contrast, the idea of an integral as being the area under a curve is much easier to state and visualize. Now we will provide an easy way to state and interpret the derivative of a curve.

Consider the function f(x) graphed in Figure (20). At a distance x down the x axis, the curve had a height f(x) as shown. Slightly farther down the x axis, at  $x + \Delta x$ , the curve has risen to a height  $f(x + \Delta x)$ .



Two points on a curve, a distance  $\Delta x$  apart.



**Figure 20a** At this point, the curve is tilted by approximately an angle  $\theta^*$ .

Figure (20a) is a blowup of the curve in the region between x and  $x + \Delta x$ . If the distance  $\Delta x$  is sufficiently small, the curve between x and  $x + \Delta x$  should be approximately a straight line and that part of the curve should be approximately the hypotenuse of the right triangle abc seen in Figure (20a). Since the side opposite to the angle  $\theta^*$  is  $f(x + \Delta x) - f(x)$ , and the adjacent side is  $\Delta x$ , we have the result that the tangent of the angle  $\theta^*$  is

$$\tan\left(\theta^*\right) = \frac{f(x + \Delta x) - f(x)}{\Delta x}$$
(156)

When we make  $\Delta x$  smaller and smaller, take the limit as  $\Delta x \rightarrow 0$ , we see that the angle  $\theta^*$  becomes more nearly equal to the angle  $\theta$  shown in Figure (21), the angle of the curve when it passes through the point x. Thus

$$\tan \theta = \lim_{\Delta x \to 0} \frac{f(x + \Delta x) - f(x)}{\Delta x}$$
(157)

The tangent of the angle at which the curve passes through the point x is called the *slope of the curve at the point x*. Thus from Equation (157) we see that the slope of the curve is equal to the derivative of the curve at that point. We now have the interpretation that the derivative of a curve at some point is equal to the slope of the curve at that point, while the integral of a curve is equal to the area under the curve up to that point.



Figure 21

The tangent of the angle  $\theta$  at which the curve passes through the point x is called the slope of the curve at that point.

# **Negative Slope**

In Figure (22) we compare the slopes of a rising and a falling curve. In (22a), where the curve is rising, the quantity  $f(x + \Delta x)$  is greater than f(x) and the derivative or slope

$$\frac{df(x)}{dx} = \lim_{\Delta x \to 0} \left| \frac{f(x + \Delta x) - f(x)}{\Delta x} \right|$$

is a positive number.

In contrast, for the downward curve of Figure (22b),  $f(x + \Delta x)$  is less than f(x) and the slope is negative. For a curve headed downward, we have

$$\frac{df(x)}{dx} = -\tan(\theta) \qquad \frac{downward heading}{curve} \quad (158)$$

(For this case you can think of  $\theta$  as a negative angle, so that  $\tan(\theta)$  would automatically come out negative. However it is easier simply to remember that the slope of an upward directed curve is positive and that of a downward directed cure is negative.)



**Figure 22** *Going uphill is a positive slope, downhill is a negative slope.* 

#### **Exercise 8**

Estimate the numerical value of the slope of the curve shown in Figure (23) at points (a), (b), (c), (d) and (e). In each case do a sketch of  $[f(x + \Delta x) - f(x)]$  for a small  $\Delta x$ , and let the slope be the ratio of  $[f(x + \Delta x) - f(x)]$  to  $\Delta x$ . Your answers should be roughly 1, 0, -1, +  $\infty$ ,  $-\infty$ .



**Figure 23** *Estimate the slope at the various points indicated.* 

## THE EXPONENTIAL DECAY

A curve that we will encounter several times during the course is the function  $e^{-ax}$  shown in Figure (24), which we call an exponential decay. Since exponents always have to be dimensionless numbers, we are writing the constant (a) in the form  $1/x_0$  so that the exponent  $x/x_0$  is more obviously dimensionless.

The function  $e^{-x/x_0}$  has several very special properties. At x = 0, it has the numerical value 1 ( $e^0 = 1$ ). When we get up to x =  $x_0$ , the curve has dropped to a value

$$e^{-x/x_0} = e^{-1} = \frac{1}{e} \quad (at \ x = x_0)$$
  
 $\approx \frac{1}{2.7}$ 
(159)

When we go out to  $x = 2x_0$ , the curve has dropped to

$$e^{-2x_0/x_0} = e^{-2} = \frac{1}{e^2}$$
 (160)

Out at  $x = 3x_0$ , the curve has dropped by another factor of e to (1/e)(1/e)(1/e). This decrease continues indefinitely. It is the characteristic feature of an exponential decay.

#### **Muon Lifetime**

In the muon lifetime experiment, we saw that the number of muons surviving decreased with time. At the end of two microseconds, more than half of the original 648 muons were still present. By 6 microsec-



#### Figure 24

As we go out an additional distance  $x_0$ , the exponential curve drops by another factor of 1/e.

onds, only 27 remained. The decay of these muons is an example of an exponential decay of the form

number of  
surviving = 
$$\begin{pmatrix} number of muons at time t = 0 \end{pmatrix} \times e^{-t/t_0}$$
 (161)

where  $t_0$  is the time it takes for the number of muons remaining to drop by a factor of 1/e = 1/2.7. That time is called the muon *lifetime*.

We can use Equation (161) to estimate the muon lifetime  $t_0$ . In the movie, the number of mesons at the top of the graph, reproduced in Figure (25), is 648. That is at time t = 0. Down at time t = 6 microseconds, the number surviving is 27. Putting these numbers into Equation (161) gives

$$27 \frac{\text{surviving}}{\text{muons}} = 648 \frac{\text{initial}}{\text{muons}} \times e^{-6/t_0}$$
$$e^{-6/t_0} = \frac{27}{648} = .042$$
(162)

Take the natural logarithm ln of both sides of Equation (162), [remembering that  $ln(e^x) = x$ ] gives

$$ln(e^{-6/t_0}) = \frac{-6}{t_0} = ln(.042) = -3.17$$

where we entered .042 on a scientific calculator and pressed the ln key. Solving for  $t_0$  we get

$$t_0 = \frac{6}{3.17} = 1.9$$
 microseconds (163)

This is close to the accepted value of  $t_0 = 2.20$  microseconds which has been determined from the study of many thousands of muon decays.



#### Figure 25

The lifetime of each detected muon is represented by the length of a vertical line. We can see that many muons live as long as 2 microseconds  $(2\mu s)$ , but few live as long as 6 microseconds.

# Half Life

The exponential decay curve  $e^{-t/t_0}$  decays to 1/e = 1/2.7 of its value at time  $t_0$ . While 1/e is a very convenient number from a mathematical point of view, it is easier to think of the time  $t_{1/2}$  it takes for half of the muons to decay. This time  $t_{1/2}$  is called the *half life* of the particle.

From Figure (26) we can see that the half life  $t_{1/2}$  is slightly shorter than the lifetime  $t_0$ . To calculate the half life from  $t_0$ , we have

$$e^{-t/t_0}\Big|_{t=t_{1/2}} = e^{-t_{1/2}/t_0} = \frac{1}{2}$$
 (164)

Again taking the natural logarithm of both sides of Equation (164) gives

$$ln\left(e^{-t_{1/2}/t_0}\right) = \frac{-t_{1/2}}{t_0} = ln\left(\frac{1}{2}\right) = -.693$$

$$\boxed{t_{1/2} = .693 t_0}$$
(165)

From Equation (165) you can see that a half life  $t_{1/2}$  is about .7 of the lifetime  $t_0$ . If the muon lifetime is 2.2µsec (we will abbreviate microseconds as µsec), and you start with a large number of muons, you would expect about half to decay in a time of

$$t_{1/2}$$
<sub>muon</sub> = .693 × 2.2µsec = 1.5 µsec

The basic feature of the exponential decay curve  $e^{-t/t_0}$  is that for every time  $t_0$  that passes, the curve decreases by another factor of 1/e. The same applies to the half life  $t_{1/2}$ . After one half life,  $e^{-t/t_0}$  has decreased to half its value. After a second half life, the curve is down to  $1/4 = 1/2 \ge 1/2$ . After 3 half lives it is down to  $1/8 = 1/2 \ge 1/2 \ge 1/2$  as shown in Figure (27).



To help illustrate the nature of exponential decays, suppose that you started with a million muons. How long would you expect to wait before there was, on the average, only one left?

To solve this problem, you would want the number  $e^{-t/t_0}$  to be down by a factor of 1 million

$$e^{-t/t_0} = 1 \times 10^{-6}$$

Taking the natural logarithm of both sides gives

$$ln(e^{-t/t_0}) = \frac{-t}{t_0} = ln(1 \times 10^{-6}) = -13.8 \quad (166)$$

(To calculate  $ln(1 \times 10^{-6})$ , enter 1, then press the *exp* key and enter 6, then press the +/– key to change it to -6. Finally press = to get the answer -13.8.)

Solving Equation (166) for t gives

$$t = 13.8 t_0 = 13.8 \times 2.2 \ \mu sec$$
  
 $t = 30 \ microseconds$  (167)

That is the nature of an exponential decay. While you have nearly half a million left after around 2 microseconds, they are essentially all gone by 30 microseconds.

#### **Exercise 9**

How many factors of 1/2 do you have to multiply together to get approximately 1/1,000,000? Multiply this number by the muon half-life to see if you get about 30 microseconds.



After each half-life, the curve decreases by another factor of 1/2.

**Figure 26** (1/2)Comparison of the lifetime  $t_0$  and the half-lifet<sub>1/2</sub>.

# Measuring the Time Constant from a Graph

The idea that the derivative of a curve is the slope of the curve, leads to an easy way to estimate a lifetime  $t_0$  from an exponential decay curve  $e^{-t/t_0}$ .

The formula for the derivative of an exponential curve is

$$\frac{\mathrm{d}\mathrm{e}^{\mathrm{a}\mathrm{t}}}{\mathrm{d}\mathrm{t}} = \mathrm{a}\mathrm{e}^{\mathrm{a}\mathrm{t}} \tag{150 repeated}$$

Setting  $a = -1/t_0$  gives

$$\frac{d}{dt}\left(e^{-t/t_0}\right) = -\frac{1}{t_0}e^{-t/t_0}$$
(168)

Since the derivative of a curve is the slope of the curve, we set the derivative equal to the tangent of the angle the curve makes with the horizontal axis.

$$\frac{d}{dt}(e^{-t/t_0}) = -\frac{1}{t_0}e^{-t/t_0} = \tan\theta$$
(168a)

The minus sign tells us that the curve is headed down.

In Figure (28), we have drawn a line tangent to the curve at the point t = T. This line intersects the (t) axis (the axis where  $e^{-t/t_0}$  goes to zero) at a distance (x) down the *t* axis.



Figure 28

A line, drawn tangent to the exponential decay curve at some point T, intersects the axis a distance x down the axis. We show that this distance x is equal to the time constant  $t_0$ . This is true no matter what point T we start with.

The height (y) of the point where we drew the tangent curve is just the value of the function  $e^{-T/t_0}$ . The tangent of the angle  $\theta$  is the opposite side (y) divided by the adjacent side (x)

$$\tan\theta = \frac{y}{x} = \frac{e^{-T/t_0}}{x}$$
(169)

Equating the two magnitudes of  $\tan \theta$  in Equations (169) in (168a) gives us

$$\frac{1}{t_0}e^{-T/t_0} = \frac{1}{x}e^{-T/t_0}$$

which requires that

$$\mathbf{x} = \mathbf{t}_0 \tag{170}$$

Equation (170) tells us that the distance (x), the distance down the axis where the tangent lines intersect the axis, is simply the time constant  $t_0$ .

The result gives us a very quick way of determining the time constant  $t_0$  of an exponential decay curve. As illustrated in Figure (29), choose any point on the curve, draw a tangent to the curve at that point and measure the distance down the axis where the tangent line intersects the axis. That distance will be the time constant  $t_0$ . We will use this technique in several laboratory exercises later in the course.



A quick way to estimate the time constant  $t_0$  for an exponential decay curve is to draw the tangent line as shown.

# THE SINE AND COSINE FUNCTIONS

The final topic in our introduction to calculus will be the functions  $\sin\theta$  and  $\cos\theta$  and their derivatives and integrals. We will need these functions when we come to rotational motion and wave motion.

The definition of  $\sin\theta$  and  $\cos\theta$ , which should be familiar from trigonometry, are

$$\sin\theta = \frac{a}{c} \left(\frac{opposite}{hypotenuse}\right)$$
 (171a)

$$\cos\theta = \frac{b}{c} \left(\frac{adjacent}{hypotenuse}\right)$$
 (171b)



where  $\theta$  is an angle of a right triangle as sown in Figure (30), (a) is the length of the side opposite to  $\theta$ , (b) the side adjacent to  $\theta$  and (c) the hypotenuse.

The formulas are simplified if we consider a right triangle whose hypotenuse is of length c = 1 as in Figure (31). Then we have

$$\sin\theta = a$$
 (172a)

 $\cos\theta = b \tag{172b}$ 



We can then fit our right triangle inside a circle of radius 1 as shown in Figure (32).



**Figure 32** *Fitting our right triangle inside a unit radius circle.* 

# **Radian Measure**

We are brought up to measure angles in degrees, but physicists and mathematicians usually measure angles in *radians*. The angle  $\theta$  measured in radians is defined as the *arc length*  $\ell$  subtended by the angle  $\theta$  on a circle of unit radius, as shown in Figure (32).

$$\theta_{\text{radians}} = \ell \qquad \begin{array}{c} arc \ length \ subtended \\ by \ \theta \ on \ a \ unit \ circle \end{array}$$
(173)

(If we had a circle of radius c, then we would define  $\theta_{radians} = \ell/c$ , a dimensionless ratio. In the special case c = 1, this reduces to  $\theta_{radians} = \ell$ .)

Since the circumference of a unit circle is  $2\pi$ , we see that  $\theta$  for a complete circle is  $2\pi$  radians, which is the same as 360 degrees. This tells us how to convert from degrees to radians. We have the conversion factor

$$\frac{360 \text{ degrees}}{2\pi \text{ radians}} = 57.3 \frac{\text{degrees}}{\text{radian}}$$
(174)

As an example of using this conversion factor, suppose we want to convert 30 degrees to radians. We would have

$$\frac{30 \text{ degrees}}{57.3 \text{ degrees/radian}} = .52 \text{ radians}$$
(175)

To decide whether to divide by or multiply by a conversion factor, use the dimensions of the conversion factor. For example, if we had multiplied 30 degrees by our conversion factor, we would have gotten

$$30 \text{ degrees} \times 57.3 \frac{\text{degrees}}{\text{radian}} = 1719 \frac{\text{degrees}^2}{\text{radian}}$$

This answer may be correct, but it is useless.

The numbers to remember in using radians are the following:

90° = 
$$\pi/2$$
 radians  
180° =  $\pi$  radians  
270° =  $3\pi/2$  radians  
360° =  $2\pi$  radians  
(176)

The other values you can work out as you need them.

# **The Sine Function**

In Figure (33) we have started with a circle of radius 1 and, in a somewhat random way, labeled 10 points around the circle. The arc length up to each of these points is equal to the angle, in radian measure, sub-tended by that point. The special values are:

$$\theta_0 = 0 \text{ radians}$$
  

$$\theta_4 = \pi/2 \text{ radians } (90^\circ)$$
  

$$\theta_6 = \pi \text{ radians } (180^\circ)$$
  

$$\theta_8 = 3\pi/2 \text{ radians } (270^\circ)$$
  

$$\theta_{10} = 2\pi \text{ radians } (360^\circ)$$

In each case the  $\sin\theta$  is equal to the height (a) at that point. For example

$$sin\theta_1 = a_1$$
  

$$sin\theta_2 = a_2$$
  

$$....$$
  

$$sin\theta_{10} = a_{10}$$

We see that the height (a) starts out at  $a_0 = 0$  for  $\theta_0$ , increases up to  $a_4 = 1$  at the top of the circle, drops back down to  $a_6 = 0$  at  $\theta_6 = \pi$ , goes negative, down to  $a_8 = -1$  at  $\theta_8 = 3\pi/2$ , and returns to  $a_{10} = 0$  at  $\theta_{10} = 2\pi$ . Our next step is to construct a graph in which  $\theta$  is shown along the horizontal axis, and we plot the value of  $\sin\theta = (a)$  on the vertical axis. The result is shown in Figure (34). The eleven points, representing the heights  $a_0$  to  $a_{10}$  at  $\theta_0$  to  $\theta_{10}$  are shown as large dots in Figure (34). We have also sketched in a smooth curve through these points, it is the curve we would get if we had plotted the value of (a) for every value of  $\theta$ from  $\theta = 0$  to  $\theta = 2\pi$ . The smooth curve is a graph of the function  $sin \theta$ .

# Exercise 10

Using the fact that the cosine function is defined as

 $\cos \theta = b$  (b is defined in Figures 31, 32)

plot the values of  $b_0, b_1, \dots, b_{10}$  on a graph similar to Figure (34), and show that the cosine function  $\cos \theta$  looks like the curve shown in Figure (35).



The heights a<sub>i</sub> at various points around a unit circle.

Figure 34 Graph of the function  $\sin(\theta)$ .

There is nothing that says we have to stop measuring the angle  $\theta$  after we have gone around once. On the second trip around,  $\theta$  increases from  $2\pi$  up to  $4\pi$ , and the curve sin $\theta$  repeats itself. If we go around several times, we get a result like that shown in Figure (36).

Several cycles of the curve  $\cos\theta$  are shown in Figure (37). You can see that the only difference between a sine and a cosine curve is where you set  $\theta = 0$ . If you move the origin of the cosine axis back (to the left) 90° ( $\pi/2$ ), you get a sine wave.



**Figure 35** *The cosine function*.

# **Amplitude of a Sine Wave**

A graph of the function  $y(\theta) = c \sin \theta$  looks just like the curve in Figure (36), except the curve goes up to a height c and down to -c as shown in Figure (38). We would get the curve of Figure (38) by plotting points around a circle as in Figure (33), but using a circle of radius c. We call this factor c the *amplitude* of the sine wave. The function  $\sin \theta$  has an amplitude 1, while the sine wave in Figure (38) has an *amplitude c* (its values range from +c to -c).





**Figure 38** A sine wave of amplitude c.







Several cycles of the curve  $\cos(\theta)$ .

# **Derivative of the Sine Function**

Since the sine and cosine functions are smooth curves, we should be able to calculate the derivatives and integrals of them. We will do this by first calculating the derivative, and then turning the process around to find the integral, just as we did for the functions  $x^n$  and  $e^x$ .

The derivative of the function  $\sin\theta$  is defined as usual by

$$\frac{d(\sin\theta)}{d\theta} = \lim_{\Delta\theta \to 0} \left[ \frac{\sin(\theta + \Delta\theta) - \sin\theta}{\Delta\theta} \right] \quad (177)$$

where  $\Delta \theta$  is a small change in the angle  $\theta$ .

The easiest way to evaluate this limit is to go back to the unit circle of Figure (25) and construct both  $\sin\theta$  and  $\sin(\theta + \Delta\theta)$  as shown in Figure (39). We see that  $\sin\theta$  is the height of the triangle with an angle  $\theta$ , while  $\sin(\theta + \Delta\theta)$  is the height of the triangle whose center angle is  $(\theta + \Delta\theta)$ . What we have to do is calculate the difference in heights of these two triangles.

In Figure (40) we start by focusing our attention on the slender triangle abc with an angle  $\Delta \theta$  at (a) and long sides of length 1 (since we have a unit circle). Since the angle  $\Delta \theta$  is small, the short side of this triangle is essentially equal to the arc length along the circle from point (b) to point (c). And since we are using radian measure, this arc length is equal to the angle  $\Delta \theta$ .



Triangles for the  $\sin\theta$  and the  $\sin(\theta + \Delta\theta)$ .

Now draw a line vertically down from point (c) and horizontally over from point (b) to form the triangle bcd shown in Figure (40). The important point is that the angle at point (c) in this tiny triangle is the same as the angle  $\theta$  at point (a). To prove this, consider the sketch in Figure (41). A line bf is drawn tangent to the circle at point (b), so that the angle abf is a right angle. That means the other two angles in the triangle add up to 90°, the total angle in any triangle being 180°

$$\theta + \varphi = 90^{\circ} \tag{178}$$

Since the angle at (e) in triangle bef is also a right angle, the other two angles in the triangle bef, must also add up to  $90^{\circ}$ .

$$\alpha + \varphi = 90^{\circ} \tag{179}$$

For both Equations (178) and (179) to be true, we must have  $\alpha = \theta$ .



The final step is to note that when  $\Delta \theta$  in Figure (40) is very small, the side cb of the very small triangle is essentially tangent to the circle, and thus parallel to the side bf in Figure (41). As a result the angle between cb and the vertical is also the same angle  $\theta$ .

Because the tiny triangle, shown again in Figure (42) has a hypotenuse  $\Delta \theta$  and a top angle  $\theta$ , the vertical side, which is equal to the difference between sin $\theta$  and sin $(\theta + \Delta \theta)$  has a height (cos  $\theta$ ) $\Delta \theta$ . Thus we have

$$\sin(\theta + \Delta \theta) - \sin \theta = (\cos \theta) \Delta \theta \tag{180}$$

Equation (180) becomes exact when  $\Delta \theta$  becomes an infinitesimal angle.

We can now evaluate the derivative

$$\frac{d(\sin\theta)}{d\theta} = \lim_{\Delta\theta\to0} \left[ \frac{\sin(\theta + \Delta\theta) - \sin\theta}{\Delta\theta} \right]$$
$$= \lim_{\Delta\theta\to0} \left[ \frac{(\cos\theta)\Delta\theta}{\Delta\theta} \right]$$
$$= \lim_{\Delta\theta\to0} \cos\theta$$

Thus we get the exact result

$$\frac{\mathrm{d}}{\mathrm{d}\theta}\left(\sin\theta\right) = \cos\theta \tag{181}$$



**Figure 42** The difference between  $\sin\theta$  and  $\sin(\theta + \Delta\theta)$  is equal to  $\Delta\theta\cos\theta$ .

#### Exercise 11

Using a similar derivation, show that

$$\frac{\mathrm{d}}{\mathrm{d}\theta}(\cos\theta) = -\sin\theta \tag{182}$$

## Exercise 12

Using the chain rule for differentiation, show that

$$\frac{d}{d\theta}(\sin a\theta) = a \cos a\theta$$

$$\frac{d}{d\theta}(\cos a\theta) = -a \sin a\theta$$

$$(a = constant) \quad (183)$$

(Hint—if you need to, look at Equation (145) through (150).

#### Exercise 13

Using the fact that integration reverses differentiation, as we did in integrating the function  $e^x$  (Equations (151) through (154), show that

$$\int_{\theta_{i}}^{\theta_{f}} (\cos a\theta) d\theta = \frac{1}{a} \sin a\theta \Big|_{\theta_{i}}^{\theta_{f}}$$

$$(184a)$$

$$(a = constant)$$

$$\int_{\theta_{i}}^{\theta_{f}} (\sin a\theta) d\theta = -\frac{1}{a} \cos a\theta \Big|_{\theta_{i}}^{\theta_{f}}$$

$$(184b)$$

Use sketches of the integrals from  $\theta_i = 0$  to  $\theta_f = \pi/2$  to show that Equations (184a) and (184b) have the correct numerical sign. (Explicitly explain the minus sign in (184b).

# P2000 Index

# **Symbols**

 $\mu$  zero, definition of 31-19  $\pi$  meson 40-23  $\Delta E\Delta t>=h$ , Uncertainty principle 40-19  $\Delta x\Delta p>=h$ , Uncertainty principle 40-15 10 dimensions, String theory int-16 100 billion to one, Matter over anti matter int-30 13.6 eV, hydrogen spectrum 35-4 1836 times, proton/electron mass ratio int-17 1987 supernova 6-14, 20-14 2.74 degrees, cosmic background radiation int-29 2D or 3D? equipotential plotting experiment 25-8 4 dimensions int-16

# A

Aberration Astigmatism Optics-21 Chromatic Optics-21 Newton's reflecting telescope Optics-22 Spherical Optics-21 In Hubble telescope mirror Optics-22 Absolute zero 17-9, 17-21 Abundance of the elements 34-24 AC voltage generator 30-21 Magnetic flux in 30-21 Accelerating field in electron gun 26-10 Acceleration Angular 12-3 Angular analogy 12-3 Calculus definition of 4-5, Cal 1-7 Component equations Cal 1-8 Vector equation Cal 1-7 Constant acceleration formulas Calculus derivation 4-9, Cal 1-20 In three dimensions 4-11, Cal 1-22 Definition of 3-13 Due to gravity 3-21 From a strobe photograph 3-15 Intuitive discussion 3-20 On inclined plane 9-11 Radial 12-5 Tangential 12-4 Uniform circular motion Direction of 3-18 Magnitude of 3-18 Vector, definition of 3-15 Acceleration versus time graphs 4-7 Accelerators, particle int-1, 28-22 Accurate values of Fourier coefficients 16-32 Adding sines and cosines in Fourier analysis 16-28 Addition of charge 19-10 Addition of forces 9-2

Adiabatic expansion Calculation of work 18-26 In Carnot cycle 18-11 Introduction to 18-9 Air cart Analysis of coupled carts 16-12 Construction of 6-2 In impulse experiments 11-9 In recoil experiments 6-2 Oscillating cart 14-5 Speed detector 30-5. See also Experiments II: - 6-Faraday's law air cart speed detector Air Resistance Calculus analysis for projectile motion 4-12 Computer analysis for projectile motion 5-24, 8-3 Strobe analysis for projectile motion 3-22 Airplane wing, Bernoulli's equation 23-13 Allowed orbits, Bohr theory int-8, 35-1 Allowed projections, spin 39-3 Allowed standing wave patterns 37-1 Alpha particles 20-8 Amount of sin(3t) present in a wave 16-28 Ampere Definition of 27-2 MKS units 24-2 Ampere's law Applied to a solenoid 29-15 Chapter on 29-1 Derivation of line integral 29-7 Field of straight wire 29-11 Final result 29-11 Maxwell's correction to 32-4 Amplitude And intensity, Fourier analysis lecture 16-33 And phase Fourier analysis lecture 16-31 Wave motion 15-17 Diffraction pattern by Fourier analysis 16-33 Fourier coefficients 16-32 Of a sine wave Cal 1-37 Analysis Fourier 16-6 Of coupled air carts 16-12 Of path 1 for electromagnetic pulse 32-14 Of path 2 for electromagnetic pulse 32-16 Analytic solution Of the RC circuit 27-22 Oscillation of mass on spring 14-7 Projectile motion with air resistance 4-12 Anderson, C., positrons int-13 Andromeda galaxy int-2, int-3, 1-22 Angle of reflection (scattering of light) 36-3, Optics-1 Angles of incidence and reflection Optics-3 Angular acceleration 12-3 Angular analogy 12-3 For Newton's second law 12-14 Torque (angular force) 12-15

#### Index-2

Angular frequency Definition of 14-4 Wave motion 15-14 Angular magnification of magnifier Optics-39 Angular mass Moment of inertia 12-7 Rotational kinetic energy 12-22 Angular momentum As a Vector 7-14, 12-7 Movie 7-15 Bohr model 35-1, 35-8 Planck's constant 35-8 Conservation of 7-9, 12-16 Derivation from F = ma 12-16 X. See also Experiments I: - 4- Conservation of angular momentum Definition of 7-10 Definition of, more general 7-12 Definition of, still more general 12-6 Definition of, cross product 12-11 Formation of planets 7-17 Gyroscopes 12-18 Kepler's second law 8-32 Magnetic moment 31-24 Movie on vector nature 7-15, 12-6, 12-17 Of bicycle wheel 12-6 Projections of, classical 7-14 Projections of electron spin 39-3 Quantized int-9 Quantized projections 38-5 Quantum number 38-7 Angular velocity As a vector 12-7 Definition of 12-2 Mass on spring 14-9 Oscillating cart 14-5 Annihilation of antimatter 34-17 Antielectron int-13 Antielectron type neutrino int-22 Antimatter 34-16 Annihilation of 34-17 Excess of matter over, in early universe 34-17, 34-29 Introduction to int-12, 34-16 Neutrino int-22 Neutron int-13 Positron int-13 Positron electron pair 34-17 Proton int-13 Wave equation for 15-2 Antiparticle int-13 Created by photon 34-17 Applications of Bernoulli's equation 23-12 Airplane wing 23-13 Aspirator 23-16 Hydrostatics 23-12 Leaky tank 23-12 Sailboat 23-14 Venturi meter 23-15

Applications of Faraday's law 30-21 AC voltage generator 30-21 Gaussmeter 30-23 Applications of Newton's Second Law 9-1 Applications of the second law of thermodynamics 18-17 Arbitrary wave, Fourier analysis 16-28 Area As a vector 24-22 Negative or positive 16-29 Related to integration Cal 1-11 Under the curve Cal 1-12 Arecibo radio telescope int-15, Optics-48 Arithmetic of vectors. See also Vector Addition 2-3 Associative law 2-4 Commutative law 2-4 Multiplication by number 2-5 Negative of 2-5 Scalar or dot product 2-12, 10-13 Subtraction of 2-5 Vector cross product 2-15, 12-9 Aspirator, Bernoulli's equation 23-16 Associative law, Exercise 2-7 Astigmatism Optics-21 Astronomy 1987 supernova int-19, 6-14, 20-14 Abundance of the elements 34-24 Big bang model of universe 33-25, 34-26 Binary stars int-2 Black dwarf star int-19 Black holes 10-29 Introduction to int-19 Blackbody radiation, color of stars 34-2 Copernicus 8-25 Crab nebula 20-16 Decoupling of light and matter 34-31 Doppler effect 33-23 Eagle nebula 7- 18, Optics-44 Early universe int-27, 34-29 Escape velocity 10-28 Evolution of the universe 34-21 Excess of matter over antimatter 34-29 Expanding universe, Hubble int-3 Formation of planets 7-17 Galaxy Andromeda int-2 Introduction to int-2 Most distant int-3 Sombrero int-2 General relativity 8-29 Globular cluster 11-2 Gravitational lensing 34-20 Helium abundance in universe 34-26 Helium core of massive star 20-15 Hubble rule for expanding universe int-3 Iron core of massive star 20-15 Kepler's laws 8-24
Astronomy Continued Light years int-2 Magnetic field of the earth 28-11 Models of the universe 34-23 Neutrino 6-14, 11-21 Neutron star And black holes 20-18 In Crab nebula 20-17 Introduction to int-19 Nuclear fusion and stellar evolution 20-12 Orion nebula 7-17 Penzias and Wilson, cosmic radiation 34-27 Powering the sun 34-23 Ptolemy, epicycle in Greek astronomy 8-25 Quantum fluctuations in space 40-25 Quasar, gravitational lens 34-20 Radio galaxy Optics-48 Radio images of variable star Optics-49 Radio telescope. See Radio telescope Radio telescope, three degree radiation int-30, 34-27 Radio telescopes Optics-48 Red shift and expanding universe int-3, 33-24, 34-21 Red supergiant star 20-15 Retrograde motion of Mars 8-24 Space travel and time dilation 1-22 Star, blackbody spectrum 34-3 Steady state model of the universe 34-25 Stellar evolution int-19 Telescopes Optics-40 Arecibo radio telescope Optics-48 Galileo's Optics-41 Hubbel Space Telescope Optics-44 Issac Newton's Optics-42 Mt. Hopkins Optics-43 Mt. Palomar Optics-43 Very Large Array, radio telescopes Optics-48 Very Long Baseline Array (VLBA) Optics-49 William Hershel's Optics-43 World's Largest Optical, Keck Optics-45 Yerkes Optics-41 Thermal equilibrium of the universe 34-28 Three degree cosmic radiation int-29, 34-27 Tycho Brahe 8-25 Van Allen radiation belts 28-32 Visible universe int-3 White dwarf star 20-15 Atmospheric pressure 17-23 Atomic And molecular forces, electric interaction 19-1 Clocks 1-21 Microscopes 17-1 Scanning Tunneling Microscope Optics-51 Processes 17-4 Spectra 33-16 Structure 19-3 Units 19-22

Atoms Angular momentum quantum number 38-7 Atomic nucleus, chapter on 20-1 Atomic processes 17-4 Avogadro's law 17-24 BASIC program, hydrogen molecule ion 19-24 Beryllium in periodic table 38-13 Bohr model int-8 Boron in periodic table 38-13 Brownian motion 17-7 Chapter on 17-1, 38-1 Classical hydrogen atom 35-2 Effective nuclear charge 38-12 Electron binding energy 19-20, 38-11 Electron energy in hydrogen molecule ion 19-21 Electron spin 38-9 Equipartition of energy 17-28 Expanded energy level diagram 38-8 Failure of classical physics 17-31 Freezing out of degrees of freedom 17-32 Heat capacity 17-26 Hydrogen molecule 19-16 Introductory view of int-16 Ionic bonding 38-15 L= 0 Patterns in hydrogen 38-4 Lithium 38-12 Model atom 37-4 Molecular and atomic processes 17-1 Molecular forces 19-15 Multi electron 38-9 Nuclear matter, chapter on 20-1 Nucleus. See also Nuclear Discovery of 11-19 Particle-wave nature of matter int-10 Pauli exclusion principle 38-9 Periodic table 38-10 Potassium to krypton 38-14 Precession of, in magnetic field 39-15 Quantized projections of angular momentum 38-5 Schrödinger's equation for hydrogen 38-2 Silicon, surface (111 plane) of Optics-51 Sodium to argon 38-13 Standing wave patterns in hydrogen 38-3 Table of 19-5 Thermal motion of 17-6 Up to neon 38-13 Xenon, photograph of 17-1 Atwood's machine 9-16 Avogadro's law 17-24 Avogadro's number, the mole 17-24

### B

Balancing weights, equilibrium 13-2 Ball Spring Pendulum. See Pendulum: Spring **Balmer** series Energy level diagram for 35-6 Formula from Bohr theory 35-5 Hydrogen spectrum 35-4 Introduction to, hydrogen star 33-19 Barometer, mercury, pressure measurement 17-22 Basic electric circuits 27-1 BASIC program. See also Computer Calculating circle 5-6 Calculational loop for satellite motion 8-19 Comment lines in 5-7 Computer time step 5-14 Conservation of angular momentum 8-32 Conservation of energy 8-35 DO LOOP 5-4 For drawing circle 5-11 For hydrogen molecule ion 19-24 For oscillating cart 14-32 For oscillatory motion 14-21, 14-30 For projectile motion 5-18, 5-19, 5-21, 8-21 For projectile motion with air resistance 5-22 For satellite motion 8-21 For spring pendulum 9-20 Kepler's first law 8-26 Kepler's second law 8-27 Kepler's third law 8-28 LET Statement 5-5 Modified gravity 8-29 Multiplication 5-6 New calculational loop 8-17 Orbit-1 program 8-21 Perihelion, precession of 8-30 Plotting a point 5-6 Plotting window 5-7 Prediction of satellite orbits 8-16 Satellite motion laboratory 8-23 Selected printing (MOD command) 5-10 Sine wave products 16-29 Unit vectors 8-18 Variable names 5-6 Bathtub vortex 23-2 Baud rate, for fiber optics Optics-14 Bell Telephone Lab, electron waves 35-12 Berkeley synchrotron 28-22 Bernoulli's equation Applications of Airplane Wing 23-13 Aspirator 23-16 Leaky tank 23-12 Sailboat 23-14 Venturi meter 23-15 Applies along a streamline 23-11 Care in applying 23-16 Derivation of 23-9 Formula for 23-11 Hydrodynamic voltage 23-17

Beryllium Binding energy of last electron 38-12 In periodic table 38-13 Beta decay And energy conservation int-21 Neutrinos 20-6 Neutrons 20-7 Protons 20-7 Recoil experiment 6-6 Beta, Hans, proton cycle, energy from sun 34-23 Beta ray int-21 Betatron 30-16 Bi-concave lens Optics-27 Bi-convex lens Optics-27 **Bicycle wheel** As a collection of masses 12-5 As a gyroscope 12-18 Right hand rule for rotation 12-11 Vector nature of angular momentum 7-14, 12-6, 12-17 Big bang model of universe int-4, 33-25, 34-26 Binary stars int-2 Binding energy Hydrogen molecule ion 19-23 Molecular forces 17-13 Nuclear 20-9 Nuclear stability 20-10 Of inner electrons 38-12 Binomial expansion 1-31, Cal 1-23 Black dwarf star int-19 Black holes And neutron stars 20-18 Critical radius for sun mass 10-30 Introduction to int-19 Stellar evolution int-20 Theory of 10-29 Blackbody radiation Electromagnetic spectrum 32-22 Photon picture of 34-22 Planck's formula 34-4 Theory of 34-2 Wein's displacement law 34-2 Blood flow, fluid dynamics 23-23 Bohr magneton Dirac wave equation 39-5 Unit of magnetic moment 39-4 Bohr model int-8 Allowed orbits 35-1 Angular momentum 35-1, 35-8 Chapter on 35-1 De Broglie explanation 35-1 Derivation of 35-8 Energy levels 35-4 Introduction to int-8 Planck's constant 35-1, 35-8 Quantum mechanics 35-1 Rydberg constant 35-9 Bohr orbits, radii of 35-7

Boltzman Constant 17-11 Formula for entropy 18-24 Bonding Covalent 19-15 Ionic 38-15 Born interpretation of particle waves 40-6 Boron Binding energy of last electron 38-12 In periodic table 38-13 Bottom quark int-24 Bragg reflection 36-4 Brahe, Tycho 8-25 Brownian motion Discussion 17-7 Movie 17-7 Bubble chambers 28-26 Bulk modulus 15-8 Button labeled ø on MacScope 16-32

## С

c (speed of light) int-2, 1-12. See also Speed of light Calculating Fourier coefficients 16-28 Calculational loop 5-17 For projectile motion 5-19, 8-17 For projectile motion with air resistance 5-24 Satellite Motion 8-19 Calculations Computer, step-by-step 5-1 Of flux 24-22 Of integrals Cal 1-11 Calculus And the uncertainty principle 4-1, Cal 1-3 Calculating integrals Cal 1-11 Calculus in physics 4-1, Cal 1-3 Chain rule Cal 1-25 Definition of acceleration 4-5, Cal 1-7 Component equations Cal 1-8 Vector equation Cal 1-7 Definition of velocity 4-3. Cal 1-5 Component equations Cal 1-8 Vector equation Cal 1-6 Derivation, electric force of charged rod 24-6 Derivation of constant acceleration formulas 4-9. Cal 1-20 In three dimensions 4-11, Cal 1-22 Limiting process 4-1, Cal 1-3, Cal 1-5 Vector equation for Cal 1-5 Line integral 29-5 Special chapter on Cal 1-3 Surface integral 29-2 Calculus in Physics 4-1 Calibration of force detector 11-10 Camera Depth of field Optics-34 Pinhole Optics-35 Single lens reflex Optics-33

Capacitance Electrical 27-16 Introduction to 27-14 Capacitor Electrolytic 27-17 Energy storage in 27-18 Examples of 27-17 In circuits. See also Circuits As circuit elements 27-20 Hydrodynamic analogy 27-14 LC circuit 31-10 Parallel connection 27-20 RC circuit 27-22 Series connection 27-21 Introduction to 27-14 Magnetic field in 32-6 Parallel plate Capacitance of 27-16 Deflection plates 26-16 Introduction to 26-14 Voltage in 26-15 Carbon Burning in oxygen 17-5 Graphite crystal, electron diffraction 36-8 Carnot cycle As thought experiment 18-4 Efficiency of Calculation of 18-28 Discussion 18-12 Formula for 18-13, 18-29 Reversible engines 18-18 Energy flow diagrams 18-15 Entropy 18-22 Introduction to 18-11 Maximally efficient engines 18-15 Refrigerator, energy flow diagrams 18-15 Reversible engines 18-13 Reversiblility 18-17 Second law of thermodynamics 18-4 Cassegrain telescope Optics-42 Cavendish experiment 8-7 Center of mass Diver movie 11-1 Dynamics of 11-4 Formula for 11-3 Gravitational force acting on 13-4 Introduction to 11-2 Center of our galaxy Optics-47 Cerenkov radiation Optics-10 CERN Electroweak theory int-26 Proton synchrotron at 28-24 CGS units Back cover-1 Classical hydrogen atom 35-2 Coulomb's law 24-2 Definition of electric charge 19-8 Chain rule Cal 1-25 Proving it (almost) Cal 1-26 Remembering it Cal 1-25

Chaos 23-1 Charge Addition of 19-10 Conservation of int-21 Density, created by Lorentz contraction 28-6 Discussion of int-6 Electric, definition of (CGS units) 19-8 Fractional (quarks) int-24, 19-15 Magnetic moment for circular orbit 31-24 On electron, Millikan oil drop experiment 26-17 Positive and negative int-6, 19-10 Quantization of electric 19-14 Surface 26-2 Unit test 24-11 Charges, static, line integral for 30-2 Charm quark int-24 Chemistry. See Atoms: Angular momentum quantum number Cholera molecule 17-2 Chromatic aberration Optics-21 Newton's reflecting telescope Optics-22 Ciliary muscle, eye Optics-31 Circuits Basic 27-1 Grounding 26-8 Inductor as a circuit element 31-7 Kirchoff's law 27-10 LC circuit Experiment 31-13 Fourier analysis of 31-31 Introduction to 31-10 Ringing like a bell 31-36 LR circuit Exponential decay 31-9 Introduction to 31-8 Neon oscillator circuit 27-29 Power in 27-9 RC circuit Exponential decay 27-23 Exponential rise 27-26 Initial slope 27-25 Introduction to 27-22 Measuring time constant 27-25 Time constant 27-24 X. See Experiments II: - 3- The RC Circuit Short 27-9 Simple 27-8 The voltage divider 27-13 Circular electric field 30-13 Line integral for 30-13 Circular motion Force causing 8-2 Particles in magnetic field 28-20 Uniform Introduction to 3-17 Magnitude of acceleration 3-18 Circular orbit, classical hydrogen atom 35-2 Circular wave patterns, superposition of 33-2 Classical hydrogen atom 35-2

Classical physics int-7 Clock Atomic clocks 1-21 Lack of simultaneity 1-32 Light pulse clock 1-14 Muon clock 1-20 Time dilation 1-22 Cluster, globular 11-2 Cochlea (inside of the ear) 16-34 Coefficient of friction 9-13 Coefficients, Fourier (Fourier analysis lecture) 16-28 Coil As a circuit element 31-7 Field of a solenoid 28-17. 29-14 Inductance of 31-5 Magnetic field of Helmholtz coils 28-19 Primary 30-26 Toroidal 31-6 In LC circuit 31-11 Torroidal 29-17 Coil, primary 30-26 Collisions Discovery of the atomic nucleus 11-19 Energy loss 11-14 Experiments on momentum conservation 7-4 Force detector 11-10 Impulse 11-9 Introduction to 11-9 Momentum conservation during 11-13 Subatomic 7-7 That conserve momentum and energy (elastic) 11-16 X. See Experiments I: - 8- Collisions Color force 19-15 Colors And Fourier analysis 16-28 Blackbody radiation 32-22 Color of stars 34-2 Electromagnetic Spectrum 32-20 Glass prism and rainbow of colors Optics-15 Comment lines, computer 5-7 Commutative law, exercise on 2-7 Compass needles, direction of magnetic field 28-12 Component sine wave, Fourier analysis 16-28 Components, vector Cal 1-7 Formula for cross product 2-17 Introduction to 2-8 Compton scattering, photon momentum 34-15

Computer. See also BASIC program BASIC. See also BASIC program Calculations Introduction to 5-2 Step-by-step 5-1 Commands Comment lines 5-7 DO LOOP 5-4 LET Statement 5-5 Multiplication notation 5-6 Selected Printing (MOD command) 5-10 Variable names 5-6 English program For projectile motion 5-16, 5-19 For satellite motion 8-19 Plot of electric fields Field plot model 25-12 In electron gun 26-13 Of various charge distributions 24-19 Plotting A point 5-6 Crosses 5-11 Window 5-7 Prediction of motion 5-12 Chapter on 5-1 Satellite orbits 8-16 Satellite with modified gravity 8-30 Program for Air resistance 5-24 Damped harmonic motion 14-34 Harmonic motion 14-30 Hydrogen molecule ion 19-24 Plotting a circle 5-2, 5-4, 5-11 Projectile motion, final one 5-21 Projectile motion, styrofoam projectile 5-28 Projectile motion with air resistance 5-22 Satellite motion 8-21 Programming, introduction to 5-4 Satellite motion calculational loop 8-19 Time Step and Initial Conditions 5-14 Computer analysis of satellite motion. See Experiments I: - 5- Computer analysis of satellite motion Computer prediction of projectile motion. See Experiments I: - 2- Computer prediction of projectile motion Computers Why they are so good at integration Cal 1-12 Conductors And electric fields, chapter on 26-1 Electric field in hollow metal sphere 26-4 Electric field inside of 26-1 Surface charge density 26-3 Cones, nerve fibers in eye Optics-31 Conical pendulum 9-18 And simple pendulum 14-17

Conservation of Angular momentum 8-32 Derivation from F = ma 12-16 Introduction to 7-9 Electric charge 19-13 Energy int-11, 8-35 Feynman's introduction to 10-2 Mass on spring 14-11 Uncertainty principle 40-24 Work Energy Theorem 10-20 X. See Experiments I: - 9- Conservation of energy Energy and momentum, elastic collisions 11-16 Linear and angular momentum, chapter on 7-1 Linear momentum 7-2, 11-7 during collisions 11-13 Conservative force 25-5 And non-conservative force 10-21 Definition of 29-6 Conserved field lines, flux tubes 24-17 Constant acceleration formulas Angular analogy 12-3 Calculus derivation 4-9, Cal 1-20 In three dimensions 4-11, Cal 1-22 Constant, integral of Cal 1-13 Constant voltage source 27-15 Continuity equation For electric fields 24-14 For fluids 23-5 Continuous creation theory int-4 Contour map 25-1 Contraction, Lorentz relativistic 1-24 Conversion factors Back cover-1 Cook's Bay, Moorea, rainbow over Optics-17 Coordinate system, right handed 2-18 Coordinate vector Definition of 3-11 In computer predictions 5-12, 8-17 In definition of velocity vector 3-13 Cornea Optics-31 Corner reflector How it works Optics-7 On the surface of the moon Optics-7 Cosine function Amplitude of Cal 1-37 Definition of Cal 1-35 Derivative of Cal 1-38 Cosine waves Derivative of 14-8 Fourier analysis lecture 16-28 Phase of 14-6 Cosmic background neutrinos int-30 Cosmic background radiation int-30, 34-27 Cosmic radiation int-30 Cosmic rays int-13

#### Index-8

Coulomb's law And Gauss' law. chapter on 24-1 Classical hydrogen atom 35-2 For hydrogen atom 24-4 For two charges 24-3 Units, CGS 24-2 Units, MKS 24-2 Coupled air cart system, analysis of 16-12 Covalent bonding 19-15 Cp = Cv + R, specific heats 18-7 Cp and Cv, specific heats 18-6 Crab nebula 20-16 Creation of antimatter, positron-electron pairs 34-17 Critical damping 14-23 Cross product Angular momentum 12-11 Component formula for 2-17 Discussion of 2-15 Magnitude of 2-17 Review of 12-9 Right hand rule 12-10 Crystal Diffraction by Thin 36-6 Graphite, electron diffraction by 36-8 Structures Graphite 36-8 Ice, snowflake 17-4 X ray diffraction 36-5 Crystalline lens, eye Optics-31 Current and voltage Fluid analogy 27-6 Ohm's law 27-7 Resistors 27-6 Current, electric Inertia of (inductance) 31-12 Introduction to 27-2 Magnetic force on 31-18 Positive and negative 27-3 Current loop Magnetic energy of 31-22 Torque on 31-20 Currents Magnetic force between 28-14, 31-19 Curve Area under, integral of Cal 1-12 Slope as derivative Cal 1-30 That increases linearly, integral of Cal 1-13 Velocity, area under Cal 1-12 Curved surfaces, reflection from Optics-3 Cycle, Carnot 18-11

### D

Damped harmonic motion Computer program for 14-34 Differential equation for 14-21 Damping, critical 14-23 Davisson & Germer, electron waves 35-12 De Broglie Electron waves int-10, 35-11 Formula for momentum 35-11 Hypothesis 35-10 Introduction to wave motion 15-1 Key to quantum mechanics 35-1 Wavelength, formula for 35-11 Waves, movie of standing wave model 35-11 Debye, on electron waves 37-1, 38-2 Decay Exponential decay Cal 1-32 Decoupling of light and matter in early universe 34-31 Definite integral Compared to indefinite integrals Cal 1-14 Defining new functions Cal 1-15 Introduction to Cal 1-11 Of velocity Cal 1-11 Process of integrating Cal 1-13 Deflection plates in electron gun 26-16 Degrees of freedom Freezing out of 17-32 Theory of 17-28 Depth of field, camera Optics-34 Derivative As a limiting process Cal 1-6, Cal 1-18, Cal 1-23, Cal 1-28, Cal 1-30 Constants come outside Cal 1-24 Negative slope Cal 1-31 Of exponential function e to the x Cal 1-28 Of exponential function e to the ax Cal 1-29 Of function x to the n'th power Cal 1-24 Of sine function Cal 1-38 Derivative as the Slope of a Curve Cal 1-30 Descartes, explanation of rainbow Optics-16 Description of motion 3-3 Detector for radiated magnetic field 32-26 Diagrams, PV (pressure, volume) 18-8 Differential equation For adiabatic expansion 18-27 For damped harmonic motion 14-21 For forced harmonic motion 14-25, 14-28 For LC circuit 31-10 For LR circuit 31-9 For oscillating mass 14-8 Introduction to 4-14 Differentiation. See also Derivative Chain rule Cal 1-25 More on Cal 1-23

Differentiation and integration As inverse operations Cal 1-18 Velocity and position Cal 1-18, Cal 1-19 Fast way to go back and forth Cal 1-20 Position as integral of velocity Cal 1-20 Velocity as derivative of position Cal 1-20 Diffraction By thin crystals 36-6 Electron diffraction tube 36-9 Of water waves 33-5 X Ray 36-4 Diffraction, electron. See Experiments II: -11- Electron diffraction experiment Diffraction grating 33-12. See also Experiments II: -10- Diffraction grating and hydrogen spectrum Diffraction limit, telescopes Optics-45 Diffraction pattern 16-33, 33-5 Analysis of 36-11 By strand of hair 36-14 Electron 36-10 For x rays 36-5 Of human hair 36-14 Recording 33-28 Single slit 33-27 Student projects 36-13 Two-slit 33-6 Dimensional analysis For predicting the speed of light 15-9 For predicting the speed of sound 15-6 Dimensions Period and frequency 14-4 Using, for remembering formulas 14-4 Dimensions of Capacitance Front cover-2 Electric charge Front cover-2 Electric potential Front cover-2 Electric resistance Front cover-2 Energy Front cover-2 Force Front cover-2 Frequency Front cover-2 Inductance Front cover-2 Magnetic field Front cover-2 Magnetic flux Front cover-2 Power Front cover-2 Pressure Front cover-2 Dirac equation Antimatter int-13, 15-2, 34-16 Electron spin 39-3 Bohr magneton 39-5 Dirac, P. A. M., prediction of antiparticles int-13 Direction an induced electric field 31-3 Direction of time And strobe photographs 3-27 Dive movie 18-1 Entropy 18-25 Neutral K meson 18-25 Rising water droplets 18-3 Discovery of the atomic nucleus 11-19

Disorder Direction of time 18-25 Entropy and the second law of thermodynamics 18-4 Formula for entropy 18-24 **Displacement vectors** From strobe photos 3-5 Introduction to 2-2 Distance, tangential 12-4 Distant galaxies, Hubble photograph int-3 Dive movie, time reversed 18-1 Diver, movie of 11-1 Diverging lenses Optics-26 DO LOOP, computer 5-4 Doppler effect Astronomer's Z factor 33-23 For light 33-22 In Astronomy 33-23 Introduction to 33-20 Relativistic formulas 33-22 Stationary source, moving observer 33-21 Universe, evolution of 34-21 Dot product Definition of 2-12 Interpretation 2-14 Work and energy 10-13 Down quark int-24 Drums, standing waves on 16-22 Duodenum, medical imaging Optics-15

### E

e - charge on an electron 19-9 E = hf, photoelectric effect formula 34-7 E = mc2, mass energy int-11, 10-3 E.dl meter 30-18 Eagle nebula Big photo of 7-18 Hubble telescope photo Optics-44 Planet formation 7-16 Ear, human Inside of cochlea 16-34 Structure of 16-15 Early universe. See Universe, early Earth Gravitational field inside of 24-24 Mass of 8-8 Earth tides 8-12 Eclipse expedition, Eddington 34-19 Edit window for Fourier transform data 16-32 Effective nuclear charge, periodic table 38-12 Efficiency Of Carnot cycle, calculation of 18-26 Of electric cars 18-19 Of heat pump 18-19 Of reversible engines 18-18

Einstein General relativity int-15, 8-29 Mass formula 6-10 Photoelectric effect int-8 Photoelectric effect formula 34-7 Principle of relativity, chapter on 1-12 Einstein cross, gravitational lens 34-20 Elastic collisions, conservation, energy, momentum 11-16 Elasticity of rubber 17-35 Electric and weak interactions unified 19-3 Electric cars, efficiency of 18-19 Electric charge Conservation of 19-13 Definition (CGS units) 19-8 Definition (MKS units) 31-19 Quantization of 19-14 Electric circuits Basic 27-1 Grounding 26-8 Kirchoff's law 27-10 LC circuit, oscillation of 31-10 LR circuit, exponential decay of 31-9 Power in 27-9 RC circuit Equations for 27-22 Exponential decay of 27-23 Half life 27-25 Initial slope 27-25 Time constant 27-24 The voltage divider 27-13 Electric current Inertia of, due to inductance 31-12 Introduction to 27-2 Positive and negative 27-3 Electric discharge of Van de Graaff generator 26-7 Electric field And conductors 26-1 And light int-7, 32-20 Circular electric field Introduction to 30-13 Line integral for 30-13 The betatron 30-16 Computer plot, -3,+5 charges 24-19 Computer plotting programs 25-12 Continuity equation for 24-14 Contour map 25-1 Created by changing magnetic flux 31-2 Created by moving magnetic field, Lorentz force 30-Direction of, when created by magnetic flux 31-3 Energy density in 27-19 Equipotential lines 25-3 Flux, definition of 24-15 Gauss' law 24-20 In electromagnetic waves 32-18 Inside a conductor 26-1 Integral of E.dl meter 30-20

Electric field continued Introduction to 24-10 Line integral of 30-14 Lines 24-12 Mapping 24-12 Mapping convention for 24-17 Of a line charge Using calculus 24-6 Using Gauss' law 24-21 Of electromagnet 30-15 Of static charges, conservative field 30-2, 30-16 Radiation by line charge 32-28 Radiation by point charge 32-30 Van de Graaff generator 26-6 Electric force Between garden peas 19-12 Produced by a line charge 24-6 Produced by a short rod 24-9 Electric force law Four basic interactions 19-2 In CGS units 19-8 Introduction to 19-7 Lorentz force law, electric and magnetic forces 28-15 Electric force or interaction int-6, int-13 Atomic & molecular forces 19-1 Electroweak theory int-26 Strength of Between garden peas 28-2 Comparison to gravity int-6, 19-8 Comparison to nuclear force int-18, 20-2 Origin of magnetic forces 28-6 Electric potential Contour map 25-1 Field plots 25-1 Of a point charge 25-5 Plotting experiment 25-7 Electric potential energy. See Potential energy: Electric Electric voltage. See also Voltage Introduction to 25-6 Van de Graaff generator 26-6 Electrical capacitance 27-16. See also Capacitor Electrically neutral int-6 Electromagnet 31-28 Electromagnetic radiation Energy radiated by classical H atom 35-3 Observed by telescopes Infrared Optics-46 Radio Optics-48 Visible Optics-42 Pulse 32-10 Analysis of path 1 32-14 Analysis of path 2 32-16 Calculation of speed 32-14 Electromagnetic spectrum int-7, 32-20, 34-11 Photon energies 34-11 Electromagnetic waves 32-18 Probability wave for photons 40-7

Electron Beam, magnetic deflection 28-9 Charge on 19-8 Diffraction Pattern 36-10 In classical hydrogen atom 35-2 Lepton family int-22 Mass in beta decay 6-7 Motion of in a magnetic field. See Experiments II: - 5-Motion of electrons in a magnetic field Radius 39-3 Rest energy in electron volts 26-12 Spin And hydrogen wave patterns 38-9 Chapter on 39-1 Spin resonance Details of experiment 39-9 Introduction to 39-5 X. See Experiments II: -12- Electron spin resonance Stability of 19-14 Two slit experiment for 40-3 Electron binding energy A classical approach 19-21 And the periodic table 38-11 In classical hydrogen molecule ion 19-23 Electron diffraction experiment 36-8 Diffraction tube 36-9 X. See Experiments II: -11- Electron diffraction experiment Electron gun Accelerating field 26-10 Electron volt 26-12 Equipotential plot 26-11, 26-13 Filament 26-9 In magnetic field Bend beam in circle 28-20 Magnetic focusing 28-29 Introduction to 26-8 X. See Experiments II: - 2- The Electron Gun Electron positron pair 34-17 Electron scattering Chapter on scattering 36-1 First experiment on wave nature 35-12 Electron screening, periodic table 38-10 Electron type neutrino int-22 Electron volt As a Unit of Energy 19-21, 26-12 Electron gun used to define 26-12 Electron waves Davisson & Germer experiment 35-12 De Broglie picture 35-11 In hydrogen 38-1 Scattering of 35-12 Wavelength of 36-9 Electroweak interaction 19-3 Theory of int-26 Weak interaction int-26 Z and W mesons int-26 Electroweak interactions 19-3

Elementary particles A confusing picture int-22 Short lived 40-23 Elements Abundance of 34-24 Creation of int-4 Table of 19-5 Ellipse Becoming a parabola Optics-4 Drawing one 8-26, Optics-3 Focus of 8-26, Optics-3 Empty space, quantum fluctuations 40-25 Energy Bernoulli's equation 23-10 Black holes 10-29 Capacitors, storage in 27-18 Chapter on 10-1 Conservation of energy And the uncertainty principle 40-24 Conservative and non-conservative forces 10-21 Derivation from work theorem 10-20 Feynman story 10-2 In collisions 11-14 In satellite motion 8-35 Mass on spring 14-11 Neutrinos in beta decay 11-20 Overview int-11 Work energy theorem 10-20 X. See Experiments I: - 9- Conservation of energy E = Mc2 10-3 Electric potential energy And molecular force 17-12 Contour map of 25-1 In classical hydrogen atom 35-3 In hvdrogen atom int-11 In hydrogen molecule ion 19-21 In nuclear fission int-18, 20-5 Negative and positive 25-4 Of a point charge 25-5 Plotting 25-7. See also Experiments II: - 1- Potential plotting Storage in capacitors 27-18 Electron binding and the periodic table 38-11 Electron, in the hydrogen molecule ion 19-21 Electron volt as a unit of energy 19-21, 26-12 Energy density in an electric field 27-19 Energy level 35-1 Energy loss during collisions 11-14 Equipartition of energy 17-28 Failure of classical physics 17-31 Freezing out of degrees of freedom 17-32 Real molecules 17-30 From nuclear fission 20-4 From sun int-18

**Energy Continued** Gravitational potential energy int-11, 8-35 Bernoulli's equation 23-10 Black holes 10-29 In a room 10-25 In satellite motion 8-36, 10-26 In stellar evolution 20-13 Introduction to 10-8 Modified 8-37 On a large scale 10-22 Zero of 10-22 Joules and Ergs 10-4 Kinetic energy int-8 Always positive 8-35 Bohr model of hydrogen 35-3 Classical hydrogen atom 35-3 Electron diffraction apparatus 36-9 Equipartition of energy 17-28 Escape velocity 10-28 Hydrogen molecule ion 19-21 Ideal gas law 17-18 In collisions 11-14 In model atom 37-5 Nuclear fusion 20-12 Origin of 10-5 Oscillating mass 14-11 Overview int-8 Pendulum 10-10 Relativistic definition of 10-5 Rotational 12-22 Satellite motion 8-36, 10-26 Slowly moving particles 10-6, 10-29 Temperature scale 17-11 Theorem on center of mass 12-26 Thermal motion 17-6 Translation and rotation 12-24 Work energy theorem 10-18 Magnetic energy of current loop 31-22 Mass energy int-18, 10-3 Negative and positive potential energy 25-4 Neutron mass energy int-22 Nuclear potential energy Fusion int-18, 20-12 Nuclear binding 20-9 Nuclear energy well. 20-10 Nuclear structure int-22 Pendulum motion, energy in 10-10 Photon energy 34-9 Photon pulse, uncertainty principle 40-21 Potential. See also Potential energy Powering the sun 34-23 Rest energy of electron and proton 26-12 Rotational kinetic energy 12-26 Spin magnetic energy Dirac equation 39-15 Magnetic moment 39-4 Magnetic potential 39-1 Spring potential energy int-11, 10-16, 14-11 Thermal energy int-8, 17-7

Total energy Classical H atom 35-3 Escape velocity 10-28 Satellite motion 8-36, 10-26 Uncertainty principle Energy conservation 40-24 Energy-time form of 40-19 Fourier transform 40-20 Voltage as energy per unit charge 25-6 Work Conservation of energy int-11 Definition of 10-12 Vector dot product 10-13 Work energy theorem 10-18 X Ray photons, energy of 36-4 Zero of potential energy 10-22 Zero point energy 37-7 Chapter on 37-1 Energy flow diagrams for reversible engines 18-15 Energy from sun, proton cycle 34-24 Energy, kinetic, in terms of momentum 37-5 Energy level diagram Balmer series 35-6 Bohr theory 35-4 Expanded 38-8 Lyman series 35-6 Model atom 37-4 Paschen series 35-6 Photon in laser 37-4 Energy-time form of the uncertainty principle 40-19 Engines Internal combustion 18-21 Maximally efficient 18-15 Reversible, efficiency of 18-18 English program For oscillatory motion 14-31 For projectile motion 5-16 For satellite motion 8-19 English program for projectile motion 5-16 Entropy Boltzman's formula for 18-24 Definition of 18-22 Number of ways to hang tools 18-23 Second law of thermodynamics 18-1 Epicycle, in Greek astronomy 8-25 Equations, differential. See Differential equation Equations, vector Components with derivatives Cal 1-7 In component form 2-10 Equilibrium Balancing weights 13-2 Chapter on 13-1 Equations for 13-2 Example - bridge problem 13-9 Example - wheel and curb 13-5 How to solve equilibrium problems 13-5 Thermal equilibrium 17-8 Working with rope 13-10

Equipartition of energy Failure of classical physics 17-31 Freezing out of degrees of freedom 17-32 Normal modes 17-28 Real molecules 17-30 Theory of 17-28 Equipotential lines 25-3 Model 25-10 Plotting experiment, 2D or 3D? 25-8 Equipotential plot for electron gun 26-11 Ergs and joules 10-4 Ergs per second, power in CGS units 24-2 Escape velocity 10-28 Euler's number e = 2.7183... Cal 1-17 Evaporation Of water 17-5 Surface tension 17-14 Even harmonics in square wave 16-28 Evolution Of stars int-19, 17-17, 20-13 Neutrinos role in 20-15 Of the universe 34-21 Excess of matter over antimatter 34-29 Exclusion principle 38-1, 38-9 Exercises, finding them. See under x in this index Expanding gas, work done by 18-5 Expanding universe Hubble, Edwin int-3 Hubble rule for int-3 Red shift 33-24, 34-19, 34-21 Expansion Adiabatic Carnot cycle 18-11 Equation for 18-26 PV Diagrams 18-9 Reversible engines 18-13 Isothermal Carnot cycle 18-11 Equation for 18-26 PV Diagrams 18-8 Reversible engines 18-13 Thermal 17-33 Uniform, of the universe int-3 Expansion, binomial 1-31, Cal 1-23 Experimental diffraction pattern 16-33 Experiments I - 1- Graphical analysis of projectile motion 3-17 - 2- Computer prediction of projectile motion 5-21 - 3- Conservation of linear momentum 7-4 - 4- Conservation of angular momentum 7-10 - 5- Computer analysis of satellite motion 8-23 - 6- Spring pendulum 9-4 - 7- Conservation of energy Check for, in all previous experiments 10-26 - 8- Collisions 11-9 - 9- The gyroscope 12-18 -10- Oscillatory motion of various kinds 14-2 -11- Normal modes of oscillation 16-4 -12- Fourier analysis of sound waves 16-18

Experiments II - 1- Potential plotting 25-7 - 2- The electron gun 26-8 - 3- The RC circuit 27- 22 - 4- The neon bulb oscillator 27-28 - 5- Motion of electrons in a magnetic field 28- 19 - 5a- Magnetic focusing, space physics 28- 30 - 6- Faraday's law air cart speed detector 30-5 - 7- Magnetic field mapping using Faraday's law 30-24 - 8- Measuring the speed of light with LC circuit 31-15 - 9- LC circuit and Fourier analysis 31-31 -10- Diffraction grating and hydrogen spectrum 33-17 -11- Electron diffraction experiment 36-8 -12- Electron spin resonance 39-9 -13- Fourier analysis and uncertainty principle 40-21 Exponential decay Cal 1-32 In LR circuits 31-9 In RC circuits 27-23 Exponential function Derivative of Cal 1-28 Exponential decay Cal 1-32 Indefinite integral of Cal 1-29 Integral of Cal 1-29 Introduction to Cal 1-16 Inverse of the logarithm Cal 1-16 Series expansion Cal 1-28 y to the x power Cal 1-16 Eye glasses experiment Optics-36 Eye, human Ciliary muscle Optics-31 Cornea Optics-31 Crystalline lens Optics-31 Farsightedness Optics-32 Focusing Optics-32 Introduction Optics-31 Iris Optics-31 Nearsightedness Optics-32 Nerve fibers Optics-31 Cones Optics-31 Rods Optics-31 Eyepiece Optics-37

# F

F = ma. See also Newton's second law Applied to Newton's law of gravity 8-5 Applied to satellite motion 8-8 For Atwood's machine 9-16 For inclined plane 9-10 For spring pendulum 9-7 For string forces 9-15 Introduction to 8-4 Vector addition of forces 9-6 f number For camera Optics-33 For parabolic mirror Optics-5 F(t) = (1)sin(t) + (1/3)sin(3t) + ...Fourier analysis of square wave 16-28 Failure of classical physics 17-31 Faraday's law AC voltage generator 30-21 Applications of 30-15 Chapter on 30-1 Derivation of 30-11 Field mapping experiment 30-24 Gaussmeter 30-23 Induced Voltage 31-4 Line integral 30-15 One form of 30-12 Right hand rule for 30-15 The betatron 30-16 Velocity detector 30-25 *Voltage transformer* 30-26 X. See Experiments II: - 6- Faraday's law air cart speed detector; Experiments II: - 7- Magnetic field mapping using Faraday's law Farsightedness Optics-32 Fermi Lab accelerator 28-23 Feynman, R. P. int-14 FFT Data button 16-32 Fiber optics Introduction to Optics-14 Medical imaging Optics-15 Field Conserved lines, fluid and electric 24-17 Electric Circular, line integral for 30-13 *Computer plot of (–3,+5) 24-19* Continuity equation for 24-14 Created by changing magnetic flux 31-2 Direction of circular or induced 31-3 Inside a conductor 26-1 Inside hollow metal sphere 26-4 Integral of - E.dl meter 30-20 Introduction to 24-10 Line integral of 30-14 Mapping convention 24-17 Mapping with lines 24-12 Of electromagnet (turned on or off) 30-15 Of line charge 24-21 Of static charges 30-2, 30-16 Radiation by line charge 32-28 Radiation by point charge 32-30 Van de Graaff generator 26-6

Field continued Electromagnetic field 32-18 Flux, introduction of concept 24-15 Gauss' law 24-20 Gravitational field Definition of 23-3 Inside the earth 24-24 Of point mass 24-23 Of spherical mass 24-24 Magnetic field Between capacitor plates 32-6 Detector, radio waves 32-26 Direction of, north pole 28-11 Gauss's law for (magnetic monopole) 32-2 In coils 28-17 In Helmholtz coils 28-18 Interaction with Spin 39-4 Introduction to 28-10 Of a solenoid 29-14 Of a toroid 29-17 Of straight wire 29-11 Surface integral 32-2 Thought experiment on radiated field 32-11 Uniform 28-16 Visualizing using compass needles 28-12 Visualizing using iron filings 28-12 Plotting experiment 25-7 Vector field Definition of 23-3 Two kinds of 30-18 Velocity field Introduction to 23-2 Of a line source 23-7 Of a point source 23-6 Field lines Computer plots, programs for 25-12 Electric Definition of 24-12 Drawing them 24-13 Three dimensional model 25-10 Field mapping Magnetic field of Helmholtz coils 30-24 Magnetic field of solenoid 30-24 Field plots and electric potential, chapter on 25-1 Filament, electron gun 26-9 First maxima of two-slit pattern 33-8 Fission, nuclear 20-3 Fitch, Val, K mesons and the direction of time 18-27 Fluctuations, quantum, in empty space 40-25 Fluid dynamics, chapter on 23-1 Fluid flow, viscous effects 23-19 Fluorescence and reflection 40-8 Flux Calculations. introduction to 24-22 Definition of 24-15 Of magnetic field 30-11 Of velocity and electric fields 24-15 Of velocity field 23-8 Tubes of flux, definition of 24-17

Flux, magnetic AC voltage generator 30-21 Definition of 30-11 Faraday's law Line integral form 30-15 Voltage form 30-12 Field mapping experiment 30-24 Gaussmeter 30-23 In the betatron 30-16 Integral E.dl meter 30-19 Magnetic field detector 32-26 Maxwell's equations 32-8 Velocity detector 30-25 Voltage transformer 30-26 Focal length For parabolic mirror Optics-5 Negative, diverging lenses Optics-26 Of a spherical surface Optics-20 Two lenses together Optics-29 Focus Eye, human Optics-32 Of a parabolic mirror Optics-4 Of an ellipse 8-26, Optics-3 Focusing, magnetic 28-29 Focusing of sound waves 8-26, Optics-3 Force 8-2 Color force 19-15 Conservative and non-conservative 10-21 Conservative forces 25-5 Electric force Classical hydrogen atom 35-2 Introduction to 19-7 Produced by a line charge 24-6 Strength of (garden peas) 28-2 Four basic forces or interactions 19-1 Introduction to force 8-2 Lorentz force law 32-8 Magnetic force Between currents 31-19 On a current 31-18 Origin of 28-10 Magnetic force law Derivation of 28-10 Vector form 28-14 Molecular force A classical analysis 19-19 Analogous to spring force 14-20 Introduction to 19-15 Potential energy for 17-12 Non linear restoring force 14-19 Nuclear force int-18 Introduction to 20-2 Range of 20-3 Particle nature of forces int-13 Pressure force 17-16 Spring force As molecular force 14-20, 17-12 Hook's law 9-3 String force Atwood's machine 9-16 Tension 9-15

Force detector 11-10 Forced harmonic motion, differential equation for 14-25, 14-28 Forces, addition of 9-2 Formation of planets 7-17 Four basic interactions int-25, 19-1 Fourier analysis Amplitude and intensity 16-33 Amplitude and phase 16-31 And repeated wave forms 16-11 Calculating Fourier coefficients 16-28 Energy-time form of the uncertainty principle 40-20 Formation of pulse from sine waves 40-27 In the human ear 16-16 Introduction to 16-6 Lecture on Fourier analysis 16-28 Normal modes and sound 16-1 Of a sine wave 16-7 Of a square wave 16-9, 16-28 Of coupled air carts, normal modes 16-12 Of LC circuit 31-31 Of slits forming a diffraction pattern 16-33 Of sound waves. See Experiments I: -12- Fourier analysis of sound waves Of violin, acoustic vs electric 16-19 X. See Experiments II: -13- Fourier analysis & the uncertainty principle Fourier coefficients Accurate values of 16-32 Calculating 16-31 Lecture on 16-28 Fourier, Jean Baptiste 16-2 Fractional charge int-24 Freezing out of degrees of freedom 17-32 Frequencies (Fourier analysis) 16-28 Frequency Angular 15-14 Of oscillation of LC circuit 31-10 Photon energy E=hf 34-7 Spacial frequency 15-14 Frequency, period, and wavelength 15-13 Friction Coefficient of 9-13 Inclined plane 9-12 Functions obtained from integration Cal 1-15 Logarithms Cal 1-15 Fusion, nuclear int-18, 20-12

### G

Galaxy Andromeda int-2 Center of our galaxy Optics-47 Introduction to int-2 Most distant int-3 Sombrero int-2 Space travel 1-22 Galileo Falling objects (Galileo Was Right!) 8-6 Inclined plane 9-10 Portrait of 9-11 Galileo's inclined plane 9-11 Galileo's telescope Optics-41 Gallbladder operation, medical image Optics-15 Gamma = Cp/Cv, specific heats 18-7 Gamma rays 32-20, 32-22 Photon energies 34-11 Wavelength of 32-20 Gamov, George, big bang theory int-4 Garden peas Electric force between 19-12 Garden peas, electric forces between 28-2 Gas constant R 17-25 Gas, expanding, work done by 18-5 Gas law, ideal 17-18 Gaudsmit and Uhlenbeck, spin 39-1 Gauss' law Electric field of line charge 24-21 For gravitational fields 24-23 For magnetic fields 32-2 Introduction to 24-20 Solving problems 24-26 Surface integral 29-3 Gauss, tesla, magnetic field dimensions 28-16 Gaussmeter 30-23 Gell-Mann Quarks int-24, 19-14 General relativity int-15, 8-29 Modified gravity 8-29 Geometrical optics Chapter on Optics-1 Definition of Optics-2 Glass prism Optics-13 Globular cluster 11-2 Gluons, strong nuclear force int-25 Graph paper For graphical analysis 3-33 For projectile motion 3-29 Graphical analysis Of instantaneous velocity 3-26 Of projectile motion 3-17 Of projectile motion with air resistance 3-22 X. See Experiments I: - 1- Graphical analysis of projectile motion

Graphite crystal Electron diffraction experiment 36-8 Electron scattering 36-1 Structure of 36-8 Grating Diffraction 33-12 Multiple slit Fourier analysis of 16-33 Interference patterns for 33-12 Three slit 16-33 Gravitational field An abstract concept 23-3 Gauss' law for 24-23 Inside the earth 24-24 Of point mass 24-23 Of spherical mass 24-24 Gravitational force. See Gravity Gravitational lens, Einstein cross 34-20 Gravitational mass 6-5 Gravitational potential energy Energy conservation int-11 Graviton int-15 Gravity int-15 Acceleration due to 3-21 And satellite motion 8-8 Black hole int-20, 10-29 Cavendish experiment 8-7 Deflection of photons 34-19 Earth tides 8-12 Einstein's general relativity int-15 Four basic interactions 19-1 Gravitational force acting at center of mass 13-4 Gravitational potential energy Bernoulli's equation 23-10 Black holes 10-29 Conservation of Energy 8-35 Energy conservation int-11 In a room 10-25 In satellite motion 8-36 Introduction to 10-8 Modified 8-37 On a large scale 10-22 Zero of 10-22 Inertial and gravitational mass 8-8 Interaction with photons 34-18 Modified, general relativity 8-29 Newton's universal law int-15, 8-5 Potential energy. See also Potential energy: Gravitational Introduction to 10-8 On a Large Scale 10-22 Quantum theory of int-16 Strength, comparison to electricity int-6, 19-8 Weakness & strength int-20 "Weighing" the Earth 8-8 Weight 8-11 Green flash Optics-17 Grounding, electrical circuits 26-8

Guitar string Sound produced by 15-22 Waves 15-20 Waves, frequency of 15-21 Gun, electron. See Electron gun Gyromagnetic ratio for electron spin 39-14 Gyroscopes Atomic scale 39-15 Movie 12-18 Precession formula 12-21 Precession of 12-19 Theory of 12-18 X. See Experiments I: - 9- The gyroscope

### Η

h bar. Planck's constant 35-9 Hair, strand of, diffraction pattern of 36-14 Half-life In exponential decay Cal 1-33 In RC circuit 27-25 Of muons (as clock) 1-20 Of muons, exponential decay Cal 1-33 Halos around sun Optics-18 Harmonic motion Computer program 14-30 Damped Computer program 14-34 Critical damping 14-23 Differential equation for 14-21 Forced Analytic solution 14-28 Differential equation for 14-25, 14-28 Harmonic oscillator 14-12 Differential equation for 14-14 Harmonic series 16-3 Harmonics and Fourier coefficients 16-28 Hays, Tobias Dive Movie, center of mass 11-1 Dive movie, time reversed 18-1 Heat capacity 17-26 Molar 17-26 Heat pump, efficiency of 18-19 Heat, specific. See Specific heat Heisenberg, Werner 4-1, Cal 1-3 Hele-Shaw cell, streamlines 23-4 Helium Abundance in early universe 34-26 And electron spin 38-9 And the Pauli exclusion principle 38-9 Binding energy of last electron 38-12 Creation of in universe int-4 Energy to ionize 38-9 In periodic table 38-11 Isotopes helium 3 and 4 int-17

Helmholtz coils 28-17, 28-18 100 turn search coil 39-12 Electron spin resonance apparatus 39-11 Field mapping experiment 30-24. See also Experiments II: - 7- Magnetic field mapping using Faraday's law Motion of electrons in 28-20, 28-29 Uniform magnetic field inside 28-17 Hertz, Heinrich, radio waves 34-1 Hexagonal array Graphite crystal and diffraction pattern 36-9 Homework exercises, finding them. See X-Ch (chapter number): Exercise number Hooke's law 9-4 In dimensional analysis 15-7 Horsehead nebula in visible & infrared light Optics-46 Hot early universe int-4 Hubbel space telescope Optics-44 Hubble, Edwin, expanding universe int-3 Hubble photograph of most distant galaxies int-3 Hubble rule for expanding universe int-3, 33-24 Hubble telescope mirror Optics-44 Spherical aberration in Optics-22 Human ear Description of 16-15 Inside of cochlea 16-34 Human Eye Optics-31 Huygens Wave nature of light 34-1, Optics-1 Huvgens' principle 33-4 Preliminary discussion of 15-2 Hydrodynamic voltage Bernoulli's equation 23-17 Resistance 27-7 Town water supply 23-18 Hydrogen atom Angular momentum quantum number 38-7 Big bang theory int-4 Binding energy of electron 38-12 Bohr theory 35-1 Classical 35-2 Coulomb's law 24-4 Expanded energy level diagram 38-8 Quantized projections of angular momentum 38-5 Solution of Schrödinger's equation 38-2 Standing wave patterns in 38-3 The L = 0 Patterns 38-4 The L ≠ 0 Patterns 38-5 Hydrogen atom, classical Failure of Newtonian mechanics 35-3 Hydrogen bomb int-18, 20-13 Hydrogen molecule Formation of 19-16 Hydrogen molecule ion Binding energy and electron clouds 19-23 Computer program for 19-24 Formation of 19-16

#### Index-18

Hydrogen nucleus int-6, 19-3 Isotopes of 19-6 Hydrogen spectrum Balmer series 33-19, 35-4 Bohr model int-8 Experiment on 33-17 Lyman series 35-6 Of star 35-4 Paschen series 35-6 X. See Experiments II: -10- Diffraction grating and hydrogen spectrum Hydrogen wave patterns Intensity at the origin 38-5 L= 0 patterns 38-4 Lowest energy ones 38-3 Schrödinger's Equation 38-2 Hydrogen-Deuterium molecule, NMR experiment 39-12 Hydrostatics, from Bernoulli's equation 23-12

## I

IBM Labs, atomic microscopes 17-1 Ice crystal 17-4 Ideal gas law 17-18 Chemist's form 17-25 Ideal gas thermometer 17-20 Absolute zero 17-21 Image Image distance Lens equation Optics-24 Negative Optics-26 In focal plane of telescope mirror Optics-5 Medical Optics-15 Gallbladder operation Optics-15 Of duodenum Optics-15 Impulse Change in momentum 11-12 Experiment on 11-9 Measurement 11-11 Inclined plane 9-10 Galileo's inclined plane, photo of 9-11 Objects rolling down 12-25 With friction 9-12 Indefinite integral Definition of Cal 1-14 Of exponential function Cal 1-29 Index of refraction Definition of Optics-9 Glass prism and rainbow of colors Optics-15 Introduction to Optics-2 Of gas of supercooled sodium atoms Optics-9 Table of some values Optics-9 Induced voltage In moving loop of wire 30-4 Line integral for 31-4 Inductance Chapter on 31-1 Derivation of formulas 31-5

Inductor As a circuit element 31-7 Definition of 31-2 Iron core 31-29 LC Circuit 31-10 LR circuit 31-8 Toroidal coil 31-6 Inertia Inertial mass 6-5 Moment of (Angular mass) 12-7 Of a massive object 31-12 Of an electric current 31-12 Infrared light Ability to penetrate interstellar dust Optics-46 Center of our galaxy Optics-47 Horsehead nebula in visible & infrared Optics-46 In the electromagnetic spectrum int-7 Paschen series, hydrogen spectra 35-6 Wavelength of 32-20 Infrared Telescopes Optics-46 Infrared camera Optics-46 IRAS satellite Optics-47 Map of the entire sky Optics-47 Mt. Hopkins 2Mass telescope Optics-46 Viewing center of our galaxy Optics-47 Initial conditions in a computer program 5-14 Initial slope in RC circuit 27-25 Inside the cochlea 16-34 Instantaneous velocity And the uncertainty principle 4-2, Cal 1-4 Calculus definition of 4-3, Cal 1-5 Definition of 3-24 From strobe photograph 3-26 Instruments Percussion 16-22 Stringed 16-18 Violin, acoustic vs electric 16-19 Wind 16-20 Integral As a sum Cal 1-10 Calculating them Cal 1-11 Definite, introduction to Cal 1-11 Formula for integrating x to n'th power Cal 1-14, Cal 1-27 Indefinite, definition of Cal 1-14 Of 1/x, the logarithm Cal 1-15 Of a constant Cal 1-13 Of a curve that increases linearly Cal 1-13 Of a velocity curve Cal 1-12 Of exponential function e to the ax Cal 1-29 Of the velocity vector Cal 1-10 As area under curve Cal 1-12 Of x to n'th power Indefinite integral Cal 1-27

Integral, line Ampere's law 29-7 Conservative force 29-6 Evaluation for solenoid 29-15 Evaluation for toroid 29-17 Faraday's law 30-15 For circular electric field 30-13 For static charges 30-2 In Maxwell's equations 32-8 Introduction to 29-5 Two kinds of fields 30-18 Integral of E.dl meter 30-18, 30-20 Integral sign Cal 1-10 Integral, surface 29-2 For magnetic fields 32-2 Formal introduction 29-2 Gauss' law 29-3 In Maxwell's equations 32-8 Two kinds of fields 30-18 Integration Equivalent to finding area Cal 1-11 Introduction to Cal 1-8 Introduction to finding areas under curves Cal 1-13 Why computers do it so well Cal 1-12 Integration and differentiation As inverse operations Cal 1-18 Velocity and position Cal 1-19 Fast way to go back and forth Cal 1-20 Position as integral of velocity Cal 1-20 Velocity as derivative of position Cal 1-20 Integration formulas Cal 1-27 Intensity And amplitude, Fourier analysis lecture 16-33 Of diffraction pattern 16-33 Of harmonics in Fourier analysis of light pulse 40-22 Of probability wave 40-22 Sound intensity, bells and decibels 16-24 Sound intensity, speaker curves 16-27 Interactions. See also the individual forces Electric int-14 Four basic 19-1 Gravitational int-15 Nuclear int-14 Photons and gravity 34-18 Weak int-14 Interactions, four basic 19-1 Interference patterns A closer look at 33-26 Introduction to 33-3 Two-slit Light waves 33-10 Probability waves 40-9 Water waves 33-6 Internal combustion engine 18-21 Internal reflection Optics-13 Interval, evaluating variables over Cal 1-10 Ionic Bonding 38-15 Iris Optics-31

Iron 56, most tightly bound nucleus int-18, 20-11 Iron core inductor 31-29 Iron core of massive star 20-15 Iron fillings, direction of magnetic field 28- 12 Iron magnets 31-26 Isothermal expansion *Calculation of work 18-26 PV Diagrams 18-8* Isotopes of nuclei int-17, 19-6 *Stability of int-22* 

#### J

Jeweler using magnifier Optics-38 Joules and Ergs 10-4

### K

K meson and direction of time int-23, 40-23 Karman vortex street 14-25 Kepler's laws Conservation of angular momentum 8-32 First law 8-26 Introduction to 8-24 Second law 8-27 Third law 8-28 Kilobaud, fiber optics communication Optics-14 Kinetic energy Always positive 8-35 Bohr model of hydrogen 35-3 Classical hydrogen atom 35-3 Electron diffraction apparatus 36-9 Equipartition of energy 17-28 Escape velocity 10-28 Hydrogen molecule ion 19-21 Ideal gas law 17-18 In collisions 11-14 In model atom 37-5 In terms of momentum 37-5 Nonrelativistic 10-6 Nuclear fusion 20-12 Origin of 10-5 Oscillating mass 14-11 Overview int-8 Pendulum 10-10 Relativistic definition of 10-5 Rotational 12-22 Satellite motion 8-36, 10-26 Slowly moving particles 10-6, 10-29 Temperature scale 17-11 Theorem on center of mass 12-26 Thermal motion 17-6 Translation and rotation 12-24 Work energy theorem 10-18 Kirchoff's law Applications of 27-11 Introduction to 27-10

### L

L = 0 Patterns, hydrogen standing waves 38-4 Lack of simultaneity 1-32 Lambda max. See Blackbody radiation: Wein's displacement law Lambda(1520), short lived elementary particle 40-23 Landé g factor 39-14 Largest scale of distance int-25 Laser Chapter on 37-1 Diffraction patterns, Fourier analysis 16-33 Pulse in gas of supercooled sodium atoms Optics-9 Standing light waves 37-2 LC circuit Experiment on 31-13 Fourier analysis of 31-31 Introduction to 31-10 Ringing like a bell 31-36 LC oscillation, intuitive picture of 31-12 Left hand rule, as mirror image Optics-6 Leibnitz 4-1. Cal 1-3 Lens Crystalline, eye Optics-31 Diverging Optics-26 Eye glasses experiment Optics-36 Eyepiece Optics-37 Introduction to theory Optics-18 Lens equation Derivation Optics-25 Introduction Optics-24 Multiple lens systems Optics-28 Negative focal length, diverging lens Optics-26 Negative image distance Optics-26 Negative object distance Optics-27 The lens equation itself Optics-25 Two lenses together Optics-29 Magnification of lenses Optics-30 Magnifier Optics-38 Jeweler using Optics-38 Magnification of Optics-39 Magnifying glass Optics-37 Multiple lens systems Optics-28 Optical properties due to slowing of light Optics-9 Simple microscope Optics-50 Spherical surface Grinding one Optics-19 Optical properties of Optics-19 Thin lens Optics-23 Two lenses together Optics-29 Zoom lens Optics-18 Lens, various types of Bi-concave Optics-27 Bi-convex Optics-27 Meniscus-concave Optics-27 Meniscus-convex Optics-27 Planar-concave Optics-27 Planar-convex Optics-27

Lenses, transmitted waves 36-3 Lensing, gravitational 34-20 Lepton family Electron int-22 Electron type neutrino int-22 Muon int-22 Muon type neutrino int-22 Tau int-22 Tau type neutrino int-22 Leptons, conservation of int-22 LET statement, computer 5-5 Lifetime Muon, exponential decay Cal 1-32 Lifting weights and muscle injuries 13-11 Light int-7 Atomic spectra 33-16 Balmer series 33-19 Blackbody radiation 34-2 Another View of 34-22 Chapter on 33-1 Decoupling from matter in early universe 34-31 Diffraction of light By thin crystals 36-6 Fourier analysis of slits 16-33 Grating for 33-12 Pattern, by strand of hair 36-14 Patterns, student projects 36-13 Doppler effect In astronomy 33-23 Introduction to 33-20 Relativistic 33-22 Electromagnetic spectrum int-7, 32-20 Photon energies 34-11 Electromagnetic waves 32-18 Gravitational lensing of 34-20 Hydrogen spectrum. See Experiments II: -10- Diffraction grating and hydrogen spectrum Balmer formula 35-5 Bohr model. See Bohr Model Lab experiment 33-17 Spectrum of star 33-19 Infra red. See Infrared light Interaction with gravity 34-18 Interference patterns for various slits 33-12 Laser pulse in gas of supercooled sodium atoms Optics-9 Lasers, chapter on 37-1 Light pulse clock 1-14 Maxwell's theory of int-7 Microwaves. See Microwaves Mirror images Optics-6 Motion through a Medium Optics-8 Particle nature of 34-1 Photoelectric effect 34-5

Photon Chapter on 34-1 Creates antiparticle 34-17 Energies in short laser pulse 40-21 In electron spin resonance experiment 39-9 Introduction to int-8 Mass 34-12 Momentum 34-13 Thermal, 2.74 degrees int-29, 34-27, 34-31 Polarization of light 32-23 Polarizers 32-25 Prism, analogy to Fourier analysis 16-28 Radiated electric fields 32-28 Radiation pressure of 34-14 Radiation pressure, red supergiant stars 20-15 Radio waves. See Radio waves Ravs Optics-1 Red shift and the expanding universe 33-24, 34-21 Reflection 36-3, Optics-1 Reflection and fluorescence 40-8 Reflection from curved surfaces Optics-4 Spectral lines, hydrogen int-7 Bohr theory 35-4 Speed in a medium Optics-8 Speed of light Electromagnetic pulse 32-14 Experiment to measure 1-9, 31-15 Same to all observers 1-12 Structure of electromagnetic wave 32-19 Thermal, 2.74 degrees 34-27 Three degree radiation 34-27 Two-slit interference pattern for 33-10 Ultraviolet. See Ultraviolet light Visible. See Visible light Visible spectrum of 33-15 Wave equation for int-7 Waves, chapter on 33-1 X ray diffraction 36-4 X rays. See X-rays Light-hour int-2 Light-minute int-2 Light-second int-2 Light-year int-2 Limiting process 4-1, Cal 1-3 Definition of derivative Cal 1-30 In calculus Cal 1-5 Introduction to derivative Cal 1-6 With strobe photographs Cal 1-2 Line charge, electric field of Calculated using calculus 24-6 Calculated using Gauss' law 24-21 Line integral. See Integral, line Linear and nonlinear wave motion 15-10 Linear momentum. See Momentum Lines, equipotential 25-3

Lithium And the Pauli exclusion principle 38-9 Atom 38-12 Binding energy of last electron 38-12 In the periodic table 38-11 Nucleus int-6 Logarithms Integral of 1/x Cal 1-15 Introduction to Cal 1-15 Inverse of exponential function Cal 1-16 Lorentz contraction 1-24 Charge density created by 28-6 Lorentz force law And Maxwell's equations 32-8 Electric and magnetic forces 28-15 Relativity experiment 30-9 Lorenz, chaos 23-1 LR circuit 31-8 Exponential decay time constant 31-9 LRC circuit, ringing like a bell 31-36 Lyman series, energy level diagram 35-6

### М

Magnetic bottle 28-31 Magnetic constant (µ zero), definition of 28-11 Magnetic energy Of current loop 31-22 Of spin 39-1 Of spin, semi classical formula 39-14 Magnetic field 28-10 Between capacitor plates 32-6 Detector 32-26 Dimensions of, tesla and gauss 28-16 Direction of Compass needles 28-12 Definition 28-11 Iron fillings 28-12 Gauss' law for 32-2 Helmholtz coils 28-17. 28-18 100 turn search coil 39-12 Electron spin resonance apparatus 39-11 Field mapping experiment 30-24 Motion of electrons in 28-20, 28-29 In electromagnetic waves 32-18 In light wave int-7 Interaction with spin 39-4 Mapping. See Experiments II: - 7- Magnetic field mapping using Faraday's law Mapping experiment with Helmholtz coils 30-24 Motion of charged particles in 28-19 Motion of electrons in. See Experiments II: - 5- Motion of electrons in a magnetic field Oersted, Hans Christian 28-12 Of a solenoid 28-17. 29-14 Of a straight wire 29-11 Of a toroid 29-17 Of permanent magnet, experiment to measure 30-25 Radiated, a thought experiment 32-11

Magnetic field Continued Right-hand rule for current 28-13 Right-hand rule for solenoids 29-14 Surface integral of 32-2 Uniform 28-16 Between Helmholtz coils 28-17 Between pole pieces 28-16 Inside coils 28-17 Magnetic flux AC voltage generator 30-21 Definition of 30-11 Faraday's law Line integral form 30-15 Voltage form 30-12 Field mapping experiment 30-24 Gaussmeter 30-23 In the betatron 30-16 Integral E.dl meter 30-19 Magnetic field detector 32-26 Maxwell's equations 32-8 Velocity detector 30-25 Voltage transformer 30-26 Magnetic focusing 28-29. See also Experiments II: - 5a- Magnetic focusing and space physics Movie 28-30 Magnetic force Between currents 31-19 Deflection of electron beam 28-9 Movie 28-9 On a current 31-18 On electrons in a wire 30-3 Origin of 28-8 Parallel currents attract 28-14 Relativity experiment (Faraday's law) 30-9 Thought experiment (on origin of) 28-7 Magnetic force law Lorentz force law, electric and magnetic forces 28-15 Magnetic force law, derivation of F = qvB 28-10Vector form 28-14 Magnetic moment And angular momentum 31-24 Bohr magneton 39-4 Definition of 31-21 Nuclear 39-6 Of charge in circular orbit 31-24 Of electron 39-4 Of neutron 39-6 Of proton 39-6 Summary of equations 31-24 Magnetic resonance Classical picture of 39-8, 39-14 Precession of atom 39-15 Electron spin resonance experiment 39-5 X. See Experiments II: -12- Electron spin resonance Magnetism Chapter on 28-1 Thought experiment to introduce 28-4

Magnets Electromagnet 31-28 Iron 31-26 Superconducting 31-30 Magnification Definition Optics-30 Negative (inverted image) Optics-30 Of Magnifier Optics-39 Of two lenses, equation for Optics-30 Magnifier Optics-38 Jeweler using Optics-38 Magnification of Optics-39 Magnifying glass Optics-30, Optics-37 Magnitude of a Vector 2-6 Map, contour 25-1 Mapping convention for electric fields 24-17 Mars, retrograde motion of 8-24 Mass Addition of 6-4 Angular Moment of inertia 12-7 Center of mass Diver movie 11-1 Dynamics of 11-4 Formula 11-3 Introduction to 11-2 Chapter on mass 6-1 Definition of mass Newton's second law 8-3 Recoil experiments 6-2 Electron mass in relativistic beta decay 6-7 Energy In nuclei int-18 Introduction to 10-3 Of neutron int-22 Gravitational force on int-6 Gravitational mass 6-5 Inertial mass 6-5 Measuring mass 6-4 Of a moving object 6-5 Of a neutrino 6-13 Of a photon 34-12 Properties of 6-3 Relativistic formula for 6-10 Relativistic mass Beta decay 6-6 Beta decay of Plutonium 246 6-8 Beta decay of Protactinium 236 6-9 Intuitive discussion 6-6 Rest mass int-11, 6-10, 10-5 Role in mechanics 8-3 Standard mass 6-3 Zero rest mass 6-11 Mass on a spring, analytic solution 14-7 Mass, oscillating, differential equation for 14-8 Mass spectrometer 28-28 Mathematical prism, Fourier analysis 16-28 Matter over antimatter, in early universe int-30, 34-17 Matter, stability of 19-14

Mauna Kea, Hawaii Optics-45 Maxima, first, of two-slit pattern 33-8 Maximally efficient engines 18-15 Maxwell's correction to Ampere's law 32-4 Maxwell's equations 32-8 Chapter on 32-1 Failure of In classical hydrogen atom 35-2 In photoelectric effect 34-6 In empty space 32-10 Probability wave for photons 40-7 Symmetry of 32-9 Maxwell's theory of light int-7, 1-9, 32-2 Maxwell's wave equation int-7, 15-1, 32-18 Measurement limitation Due to photon momentum 40-11 Due to uncertainty principle 4-2, Cal 1-4 Two slit thought experiment 40-9 Using waves 40-10 Measuring short times using uncertainty principle 40-22 Measuring time constant from graph Cal 1-34 Mechanics Newtonian Chapter on 8-1 Classical H atom 35-3 Newton's second law 8-4 Newton's Third Law 11-6 Photon mechanics 34-12 Relativistic int-12 The role of mass in 8-3 Medical imaging Optics-15 Megabit, fiber optics communication Optics-14 Meiners, Harry, electron scattering apparatus 36-1 Meniscus-concave lens Optics-27 Meniscus-convex lens Optics-27 Mercury barometer, pressure measurement 17-22 Meter, definition of int-2 Microscope int-1, Optics-50 Atomic 17-1 Scanning tunneling microscope Optics-51 Surface (111 plane) of a silicon Optics-51 Simple microscope Optics-50 Microwaves Electromagnetic spectrum int-7, 32-20 Microwave polarizer 32-24 Photon energies 34-11 Milky Way int-2 Center of our galaxy Optics-47 Millikan oil drop experiment 26-17 Mirror images General discussion Optics-6 Reversing front to back Optics-7 Right-hand rule Optics-6 Mirror, parabolic Focusing properties of Optics-4, Optics-42 MKS units Ampere, volt, watt 24-2 Coulomb's law in 24-2

MOD command, computer 5-10 Model Atom 37-4 Chapter on 37-1 Energy levels in 37-4 Model showing equipotential and field lines 25-10 Modulus Bulk 15-8 Definition of 15-8 Young's 15-8 Molar heat capacity 17-26 Molar specific heat of helium gas 17-27 Mole Avogadro's number 17-24 Volume of 17-25 Molecular forces int-6, 14-20, 17-12 A classical analysis 19-18 The bonding region 19-19 A more quantitative look 19-18 Binding energy 17-13 Electric interaction 19-1 Four basic interactions 19-2 Introduction to 19-15 Represented by springs 17-13 Molecular weight 17-24 Molecules 17-2 Cholera 17-2 Hydrogen, electric forces in 19-16 Hydrogen molecule ion 19-16 Myoglobin 17-3 Water 17-2 Moment, magnetic And angular momentum 31-24 Definition of 31-21 Of charge in circular orbit 31-24 Summary of equations 31-24 Moment of inertia Angular mass 12-7 Calculating 12-8 Rotational kinetic energy 12-22 Momentum Angular. See Angular momentum Collisions and impulse 11-9 Conservation of 7-2 Derivation from Newton's second law 11-7 During collisions 11-13 General discussion 7-1 In collision experiments 7-4 In subatomic collisions 7-7 X. See also Experiments I: - 3- Conservation of linear momentum De Broglie formula for momentum 35-11 Kinetic energy in terms of momentum 37-5 Linear momentum, chapter on 7-1 Momentum of photon Compton scattering. 34-15 Formula for 34-13 Momentum version of Newton's second law 11-8 Uncertainty principle, position-momentum form 40-15

Mormon Tabernacle, ellipse Optics-3 Mormon Tabernacle, focusing of sound waves 8-26, Optics-3 Motion Angular analogy 12-3 Damped harmonic motion 14-21 Differential equation for 14-21 Description of, chapter on 3-3 Forced harmonic motion Differential equation for 14-25, 14-28 Harmonic motion 14-12 Computer program for 14-30 Of charged particles In magnetic fields 28-19 In radiation belts 28-32 Of electrons in a magnetic field. See Experiments II: -5- Motion of electrons in a magnetic field Of light through a medium Optics-8 Oscillatory motion 14-2. See also Experiments I: -10- Oscillatory motion of various kinds Prediction of motion 5-12 Projectile. See Projectile Motion Resonance 14-24 Tacoma Narrows bridge 14-24 Rotational motion 12-1 Angular acceleration 12-3 Angular velocity 12-2 Radian measure 12-2 Satellite. See Satellite motion Thermal motion 17-6 Translation and rotation 12-24 Uniform circular motion 3-17, 8-2 Particles in magnetic field 28-20 Wave motion, amplitude and phase 15-17 Movie Angular momentum as a Vector 7-15 Brownian motion 17-7 Circular motion of particles in magnetic field 28-20 Diver 11-1 Magnetic deflection 28-9 Magnetic focusing 28-30 Muon Lifetime 1-21 Standing De Broglie like waves 35-11 Time reversed dive, second law of thermodynamics 18-1 Mt. Hopkins telescope Optics-43 Mt. Palomar telescope Optics-43 Mu ( $\mu$ ) zero, definition of 31-19 Multi Electron Atoms 38-9 Multiple lens systems Optics-28 Multiple slit grating 16-33 Multiple slit interference patterns 33-12 Multiplication notation, computer 5-6 Multiplication of vectors By a number 2-5 Scalar or dot product 2-12 Vector cross product 2-15, 12-9

Muon And Mt. Washington, Lorentz contraction 1-29 Discovery of int-23 Half life used as clock 1-20 Lepton family int-22 Lifetime, exponential decay Cal 1-32 Movie on lifetime 1-21 Cerenkov radiation Optics-10 Muon type neutrino int-21 Muscle, ciliary, in the eye Optics-31 Muscle injuries lifting weights 13-11 Myoglobin molecule 17-3 Myopia, nearsightedness Optics-32

### N

Nature's speed limit int-12, 6-11 Nearsightedness Optics-32 Nebula Crab, neutron star 20-16 Eagle 7-16, 7-18 Orion 7-17 Negative and positive charge 19-10 Negative focal length, diverging lenses Optics-26 Negative image distance Optics-26 Negative object distance Optics-27 Negative slope Cal 1-31 Neon bulb oscillator 27-28 Experimental setup 27-31 X. See Experiments II: - 4- The Neon Bulb Oscillator Neon, up to, periodic table 38-13 Nerve fibers, human eye Optics-31 Cones Optics-31 Rods Optics-31 Net area (Fourier analysis) 16-29 Neutrino astronomy 6-14, 11-21 Neutrinos 6-13, 11-20 1987 Supernova 6-14 Beta (B) decay reactions 20-6 Cosmic background int-30 Created in the weak interaction 20-6 Electron type int-21 From the sun 6-13, 11-21 In nuclear structure 20-7 In stellar evolution 20-15 In supernova explosions 20-16 Muon type int-21 Passing through matter int-22 Pauli's prediction of int-21, 20-6 Rest mass 20-6 Stability of 19-14 Tau type int-21

Neutron Decav Beta decay reactions 20-7 Charge conservation in int-21 Energy problems int-21 Weak interaction 20-6 Formation of deuterium in early universe 34-30 In alpha particles 20-8 In isotopes int-17, 19-6 In nuclear matter 20-1 Neutron proton balance in early universe 34-30 Neutron proton mass difference in early universe 34-30 Nuclear binding energies 20-9 Quark structure of int-24 Rest mass energy int-22, 20-9 Role in nuclear structure 20-7 Neutron star And black holes int-20, 20-18 Binarv int-15 In Crab nebula 20-17 Pulsars 20-17 Stellar evolution int-19 New functions, obtained from integration Cal 1-15 Newton Particle nature of light 34-1, Optics-1 Newtonian mechanics 8-1 Classical H atom 35-3 Failure of In specific heats 17-31 In the classical hydrogen atom 35-3 The role of mass 8-3 Newton's laws Chapter on 8-1 Classical physics int-7 Gravity int-15, 8-5 Second law 8-4 And Newton's law of gravity 8-5 Angular analogy 12-14 Applications of, chapter on 9-1 Atwood's machine 9-16 Inclined plane 9-10 Momentum version of 11-8 Satellite motion 8-8 String forces 9-15 Vector addition of forces 9-6 Third law 11-6 Newton's reflecting telescope Optics-22, Optics-42 NMR experiment, the hydrogen-deuterium molecule 39-12 Nonlinear restoring forces 14-19 Nonrelativistic wave equation int-12, 15-2, 34-16 Normal modes Degrees of freedom 17-29 Fourier analysis of coupled air cart system 16-12 Modes of oscillation 16-4 X. See Experiments I: -11- Normal modes of oscillation

Nuclear Binding energy 20-9 Charge, effective, in periodic table 38-12 Energy well, binding energies 20-10 Fission int-18, 20-3 Energy from 20-4 Force int-18, 20-2 Four basic interactions 19-2 Meson, Yukawa theory int-22 Range of 20-3 Range vs. electric force int-18 Fusion int-18, 20-12 Binding energies, stellar evolution 20-13 Nuclear interaction int-14 Alpha particles 20-8 Binding energy 20-9 Neutron stars 20-17 Nuclear magnetic moment 39-6 Nuclear matter, chapter on 20-1 Nuclear reactions, element creation int-4 Nuclear stability, binding energy 20-10 Nuclear structure int-22, 20-7 Nucleon int-17 Nucleus int-17 Discovery of, Rutherford 35-1 Hydrogen int-6 Isotopes of 19-6 Large int-22 Lithium int-6 Most tightly bound (Iron 56) 20-11

### 0

Odd harmonics in a square wave 16-28 Odor of violets 17-6 Oersted, Hans Christian 28-12 Off axis rays, parabolic mirror Optics-4 Ohm's law 27-7 One cycle of a square wave 16-28 One dimensional wave motion 15-1 **Optical properties** Parabolic reflectors Optics-4 Spherical surface Optics-19 Optics, fiber Introduction to Optics-14 Medical imaging Optics-15 Optics, geometrical Chapter on Optics-1 Definition of Optics-2 Optics of a simple microscope Optics-50 Orbitals 19-15. See also Hydrogen atom: Standing wave patterns in Orbits Allowed, Bohr theory int-8 Bohr, radii of 35-7 Classical hydrogen atom 35-2 Orbit-1 program 8-21 Precession of, general relativity 8-30

Satellite. See Satellite motion

Order and disorder Direction of time 18-25 Entropy and the second law of thermodynamics 18-4 Formula for entropy 18-24 Orion nebula 7-17 Oscillation 14-1 Critical damping 14-23 Damped Computer approach 14-21 Differential equation for 14-21 LC circuit 31-10 Intuitive picture of 31-12 Mass on a spring Analytic solution 14-7 Computer solution 14-30 Differential equation for 14-8 Non linear restoring forces 14-19 Normal modes 16-4. See also Experiments I: -11-Normal modes of oscillation Period of 14-4 Phase of 14-6 Resonance Analytic solution for 14-26, 14-28 Differential equation for 14-25 Introduction to 14-24 Small oscillation Molecular forces 14-20 Simple pendulum 14-16 Torsion pendulum 14-12 Transients 14-27 Oscillator Harmonic oscillator 14-12 Differential equation for 14-14 Forced 14-28 Neon bulb oscillator 27-28 Period of oscillation 27-30 X. See also Experiments II: - 4- The neon bulb oscillator Oscillatory motion 14-2. See also Experiments I: -10- Oscillatory motion of various kinds Osmotic pressure 17-34 Overview of physics int-1

# P

Parabola How to make one Optics-4 Parabolic mirror f number Optics-5 Focusing properties of Optics-4, Optics-42 Off axis rays Optics-4 Parallel currents attract 28-14 Parallel plate capacitor 26-14 Capacitance of 27-16 voltage in 26-15 Parallel resistors 27-12 Particle Point size int-14 Systems of particles 11-1

Particle accelerators 28-22 Particle decays and four basic interactions int-23 Particle nature of light 34-1 Photoelectric effect 34-5 Particle-wave nature Born's interpretation 40-6 De Broglie picture int-10, 35-10 Energy level diagrams resulting from 37-4 Of electromagnetic spectrum 34-11 Of electrons Davisson and Germer experiment 35-12 De Broglie picture int-10, 35-10 Electron diffraction experiment 36-8 Electron waves in hydrogen 38-2 Pauli exclusion principle 38-9 Of forces int-13 Of light Electromagnetic spectrum 34-11 Photoelectric effect 34-5 Photon mass 34-12 Photon momentum 34-13 Photon waves 40-6 Photons, chapter on 34-1 Of matter int-10, 34-11 Probability interpretation of 40-6 Fourier harmonics in a laser pulse 40-22 Reflection and fluorescence 40-8 Quantum mechanics, chapter on 40-1 Two slit experiment from a particle point of view Probability interpretation 40-9 The experiment 40-3 Uncertainty principle 40-14 Energy conservation 40-24 Position-momentum form of 40-15 Quantum fluctuations 40-25 Time-energy form of 40-19 Paschen series Energy level diagram 35-6 Hydrogen spectra 35-6 Pauli exclusion principle 38-1, 38-9 Pauli, W., neutrinos int-21, 20-6 Peebles, radiation from early universe 34-27 Pendulum Conical 9-18 Energy conservation 10-10 Simple and conical 14-17 Simple pendulum 14-15 Spring pendulum 9-4 Ball spring program 9-20 Computer analysis of 9-8 F = ma 9-7 X. See Experiments I: - 6- Spring pendulum Torsion pendulum 14-12 Differential equation for 14-14 Penzias and Wilson, cosmic background radiation int-29. 34-27 Percussion instruments 16-22 Period of oscillation 14-4 Neon bulb circuit 27-30

Period, wavelength, and frequency 15-13 Periodic table 38-1, 38-10 Beryllium 38-13 Boron 38-13 Effective nuclear charge 38-12 Electron binding energies 38-11 Electron screening 38-10 Lithium 38-12 Potassium to krypton 38-14 Sodium to argon 38-13 Summary 38-14 Phase and amplitude Fourier analysis lecture 16-31 Wave motion 15-17 Phase of an oscillation 14-6 Phase transition, electroweak theory int-26 Phases of Fourier coefficients 16-32 Photoelectric effect int-8 Einstein's formula 34-7 Introduction to 34-5 Maxwell's theory, failure of 34-6 Planck's constant 34-8 Photon int-8 Blackbody radiation 34-22 Chapter on 34-1 Creates antiparticle 34-17 Electric interaction int-23 Energies in short laser pulse 40-21 Energy 34-9 Energy levels in laser 37-4 Uncertainty principle, Fourier transform 40-20 Gravitational deflection of 34-19 Hydrogen spectrum 35-5 In electron spin resonance experiment 39-9 Interaction with gravity 34-18 Mass 34-12 Mechanics 34-12 Momentum 34-13 Compton scattering 34-15 Measurement limitation 40-11 Probability wave 40-7 Rest mass int-12 Stability of 19-14 Standing waves 37-3 Thermal, 2.74 degrees int-29, 34-27, 34-31 Photon pulse Photon energy in 40-21 Probability Interpretation of 40-22 Photon waves, probability interpretation 40-6 Physical constants In CGS units Back cover-1 Pi mesons int-23 Piela, electron clouds and binding energy 19-23 Pinhole camera Optics-35 Planar-concave lens Optics-27 Planar-convex lens Optics-27 Planck, M., blackbody radiation law 34-4

Planck's constant And blackbody radiation 34-4, 34-22 Angular momentum, Bohr model 35-8 Bohr theory 35-1 In Bohr magneton formula 39-4 In de Broglie wavelength formula 35-11 In photon mass formula 34-12 In photon momentum formula 34-13 In the photoelectric effect 34-7 In the uncertainty principle 40-15, 40-19 Introduction to 34-8 Spin angular momentum 39-3 Plane, inclined 9-10 Planetary units 8-14 Planets Formation of 7-17 Plates, electron gun deflection 26-16 Plotting A point by computer 5-6 Experiment, electric potential 25-7 Potentials and fields. See Experiments II: - 1- Potential plotting Window, computer 5-7 Plutonium 246 6-8 Point Mass, gravitational field of 24-23 Particle int-14 Source, velocity field of 23-6 Polarization of light waves 32-23 Polarizer Liaht 32-25 Microwave 32-24 Polaroid, light polarizer 32-25 Position measurement, uncertainty principle 40-15 Positive and negative Charge 19-10 Electric current 27-3 Positive area in Fourier analysis 16-29 Positron (antimatter) int-13, 34-17 Positronium, annihilation into photons 34-17 Potassium to krypton, periodic table 38-14 Potential, electric Contour map 25-1 Of a point charge 25-5

Potential energy Conservative forces 25-5 Electric potential energy Contour map of 25-1 In classical hydrogen atom 35-3 In hydrogen atom int-11 In hydrogen molecule ion 19-21 In molecules 17-12 In nuclear fission int-18, 20-5 Negative and positive 25-4 Of a point charge 25-5 Potential plotting 25-7 Storage in capacitors 27-18 Energy conservation int-11, 8-35, 10-20 Conservative and non-conservative forces 10-21 Mass on spring 14-11 Uncertainty principle 40-24 Equipartition of energy 17-28 Gravitational potential energy int-11, 8-35 Bernoulli's equation 23-10 Black holes 10-29 In a room 10-25 In satellite motion 8-36, 10-26 In stellar evolution 20-13 Introduction to 10-8 Modified 8-37 On a large scale 10-22 Zero of 10-22 In collisions 11-14 Negative and positive 25-4 Nuclear potential energy int-18 Binding energies 20-9 Fusion 20-12 Spin potential energy Magnetic 39-1 Magnetic moment 39-4 Spring potential energy int-11, 10-16, 14-11 Work int-11, 10-15 Work energy theorem 10-18 Potential plotting. See Experiments II: - 1- Potential plotting Power 1 horsepower = 746 watts 18-20 Definition of watt 10-31 Efficiency of a power plant 18-18 In electric circuits 27-9 Of the sun 34-23 Sound intensity 16-24 Powers of 10, names of Front cover-2 Practical system of units 10-31 Precession Of atom, magnetic interaction 39-15 Of orbit, modified gravity 8-30 Prediction of motion Using a computer 5-12 Using calculus Cal 1-9

Pressure Atmospheric 17-23 Bernoulli's equation 23-9, 23-11 Airplane wing 23-13 Care in applying 23-16 Sailboats 23-14 Superfluid helium 23-17 Ideal gas law 17-18 In stellar evolution 17-17 Measurement, using mercury barometer 17-22 Osmotic pressure 17-34 Pressure in fluids Aspirator 23-16 Definition of 23-10 Hydrodynamic voltage 23-17 Hydrostatics 23-12 Venturi meter 23-15 Viscous effects 23-19 Pressure of a gas 17-16 Pressure of light Nichols and Hull experiment 34-14 Red giant stars 20-15, 34-15 PV diagrams 18-8 Adiabatic expansion 18-9, 18-26 Internal combustion engine 18-21 Isothermal expansion 18-8, 18-26 Reversible engines 18-13 The Carnot cycle 18-11, 18-26, 18-28 Work dome by pressure, Bernoulli's equation 23-10 Primary coil 30-26 Principle of relativity. See also Relativistic physics A statement of 1-4 And the speed of light 1-11 As a basic law of physics 1-4 Chapter on 1-1 Einstein's theory of 1-12 Introduction to 1-2 Special theory of 1-13 A consistent theory 1-32 Causality 1-36 Lack of simultaneity 1-32 Light pulse clock 1-14 Lorentz contraction 1-24 Mass energy 10-3 Nature's speed limit 6-11 Origin of magnetic forces 28-8 Photon mass 34-12 Photon momentum 34-13 Relativistic energy and momenta 28-24 Relativistic mass 6-6 Relativity experiment leading to Faraday's law 30-9 Time dilation 1-16, 1-22 Zero rest mass particles 6-11 Principle of superposition For 1 dimensional waves 15-11 For 2 dimensional waves 33-2 Preliminary discussion of 15-2, 33-1

Prism Atmospheric, the green flash Optics-17 Glass Optics-13 Glass, rainbow of colors Optics-15 Prism, mathematical (Fourier analysis) 16-28 Probability interpretation And the uncertainty principle 40-14 Of electron diffraction pattern 40-6 Of particle waves 40-6 Of photon pulse 40-22 Of two slit experiment 40-2 Probability wave For photons 40-7 Intensity of 40-22 Reflection and Fluorescence 40-8 Probe for scanning tunneling microscope Optics-51 Problem solving. See Solving problems Gauss' law problems 24-26 How to go about it 24-29 Projectile motion problems 4-16 Program, BASIC. See BASIC program Program, English For oscillatory motion 14-31 For projectile motion 5-16 For satellite motion 8-19 Project suggestion on wave speed 15-8 Projectile motion Analysis of Calculus 4-9 Computer 5-16 Graphical 3-16 And the uncertainty principle 4-2, Cal 1-4 BASIC program for 5-19 Calculus definition of velocity 4-3, Cal 1-5 Computer program for 8-21 Constant acceleration formulas Calculus derivation 4-9 Graphical analysis of 3-26 Determining acceleration for 3-16 English program for 5-16 Graph paper tear out pages 3-29 Gravitational force 8-2 Instantaneous velocity 3-24 Solving problems 4-16 Strobe photograph of 3-7 Styrofoam projectile 5-28 With air resistance Calculus analysis 4-12 Computer calculation 5-22 Graphical analysis 3-22 X. See Experiments I: - 2- Computer prediction of projectile motion Projections of angular momentum Classical 7-14 Electron spin 39-3 Quantized 38-5 Protactinium 236, recoil definition of mass 6-9

Proton int-17 In alpha particles 20-8 In atomic structure 19-4 In hydrogen molecule 19-16 In nuclear structure 20-2 In the weak interaction 20-6 Quark structure int-24 Rest energy in electron volts 26-12 Stability of 19-14 Proton cycle, energy from sun 34-24 Proton synchrotron at CERN 28-24 Proton-neutron mass difference, early universe 34-30 Ptolemy, epicycle, in Greek astronomy 8-25 Pulleys Atwood's machine 9-16 Working with 9-16 Pulsars, neutron stars 20-17 Pulse Formation from sine waves 40-27 Of electromagnetic radiation 32-10 PV = NRT 17-25 **PV** diagrams Adiabatic expansion 18-9 Carnot cycle 18-11 Isothermal expansion 18-8

# Q

Quantization of electric charge 19-14 Quantized angular momentum int-9 Angular momentum quantum number 38-7 Electron spin 38-9 Chapter on 39-1 Concept of spin 39-3 In Bohr theory 35-9 In de Broglie's hypothesis 35-10 In hydrogen wave patterns 38-3 Quantized projections 38-5 Quantized vortices in superfluids 23-22 Quantum electrodynamics int-14 Quantum fluctuations in empty space 40-25 Quantum mechanics. See also Particle-wave nature: Schrödinger's equation Bohr theory of hydrogen 35-1 Chapter on 40-1 Concept of velocity 4-2, Cal 1-4 Electron and nuclear spin 39-1 Model atom 37-4 Schrodinger's equation applied to atoms 38-1 Uncertainty principle 40-14 Zero point energy 37-7 Quantum number, angular momentum 38-7 Quantum theory of gravity int-16, int-20 Quark confinement 19-15 Quarks int-24 Quantization of electric charge 19-14 Quasars Gravitational lens, Einstein cross 34-20 Size of 34-19

# R

R, gas constant 17-25 Radar waves Photon energies in 34-11 Wavelength of 32-20 Radial acceleration 12-5 Radian measure 12-2, Cal 1-35 Radiated electromagnetic pulse 32-10 Radiated magnetic field thought experiment 32-11 Radiation Blackbody 32-22 Photon picture of 34-22 Theory of 34-2 Wein's displacement law 34-2 Cerenkov Optics-10 Electromagnetic field Analysis of path 1 32-14 Analysis of path 2 32-16 Calculation of speed 32-14 Spectrum of 32-20 Radiated electric fields 32-28 Radiated energy and the classical H atom 35-3 Radiated field of point charge 32-30 Three degree cosmic radiation int-30, 34-27 UV, X Rays, and Gamma Rays 32-22 Radiation belts, Van Allen 28-32 Radiation pressure In red supergiant stars 20-15 Of light 34-14 Radio galaxy Optics-48 Radio images of variable star Optics-49 Radio telescope int-15 Arecibo int-15 Three degree radiation 34-27 Radio telescopes Optics-48 Arecibo Optics-48 Radio galaxy image Optics-48 Radio images of variable star Optics-49 Very Large Array Optics-48 Very Long Baseline Array Optics-49 Radio waves Hertz, Heinrich 34-1 In the electromagnetic spectrum int-7, 32-20 Photon energies 34-11 Predicted from the classical hydrogen atom 35-2 Wavelength of 32-20 Radius of electron 39-3 Rainbow Glass prism Optics-15 Photograph of Optics-16 Range of nuclear force 20-3 RC circuit 27-22 Exponential decay 27-23 Exponential rise 27-26 Half-lives 27-25 Initial slope 27-25 Measuring time constant 27-25 X. See Experiments II: - 3- The RC Circuit

Reactive metal, lithium int-6, 38-9 Recoil experiments, definition of mass 6-2 Red shift and the expanding universe Doppler effect 34-21 Evolution of universe 34-21 Uniform expansion 33-24 Red supergiant star 20-15 Reflecting telescope Optics-42 Cassegrain design Optics-42 Diffraction limit Optics-45 Hubbel space telescope Optics-44 Keck, world's largest optical Optics-45 Mt. Hopkins Optics-43 Mt. Palomar Optics-43 Newton's Optics-22 Newton's Optics-42 Secondary mirror Optics-42 William Hershel's Optics-43 Reflection And fluorescence, probability interpretation 40-8 Bragg reflection 36-4 From curved surfaces Optics-3 Internal Optics-13 Of light 36-3, Optics-1 Refracting telescopes Optics-40 Galileo's Optics-41 Yerkes Optics-41 Refraction, index of Definition of Optics-9 Glass prism and rainbow of colors Optics-15 Introduction to Optics-2 Of gas of supercooled sodium atoms Optics-9 Table of some values Optics-9 Relativistic mass. See Relativistic physics: Relativistic mass Relativistic physics. See also Principle of Relativity A consistent theory 1-32 Antimatter int-12, 34-16 Black holes 10-29 Blackbody radiation 34-22 Causality 1-36 Chapter on 1-1 Clock Light pulse 1-14 Moving 1-13 Muon 1-20 Muon lifetime movie 1-21 Other kinds 1-18 Real ones 1-20 Creation of positron-electron pair 34-17 Definition of mass 6-2 Doppler effect for light 33-22 Einstein mass formula 6-10 Electric or magnetic field: depends on viewpoint 30-10 Electric or magnetic force: depends on viewpoint 28-10 Electromagnetic radiation, structure of 32-19

Relativistic physics Continued Electron mass in beta decav 6-7 Gravitational lensing 34-20 Interaction of photons and gravity 34-18 Kinetic energy 10-5 Slowly moving particles 10-6 Lack of simultaneity 1-32 Longer seconds 1-16 Lorentz contraction 1-24 Origin of magnetic forces 28-8 Thought experiment on currents 28-4 Lorentz force law 28-15 Mass energy 10-3 Mass-energy relationship int-11 Maxwell's equations 32-8 Motion of charged particles in magnetic fields 28-19 Muon lifetime movie 1-21 Nature's speed limit int-2, int-12, 6-11 Neutrino astronomv 6-14 Neutrinos 6-13 Origin of magnetic forces 28-8 Particle accelerators 28-22 Photon mass 34-12 Photon rest mass 6-12 Photon momentum 34-13 Compton scattering 34-15 Principle of relativity 1-2 As a basic law 1-4 Radiated electric fields 32-28 Red shift and the expansion of the universe 34-21 Relativistic calculations 1-28 Approximation formulas 1-30 Muons and Mt. Washington 1-29 Slow speeds 1-29 Relativistic energy and momenta 28-24 Relativistic mass 6-6 Formula for 6-10 In beta decay 6-6 Relativistic mechanics int-12 Relativistic speed limit 6-11 Relativistic wave equation int-12 Relativity experiment for Faraday's law 30-9 Short lived elementary particles 40-23 Space travel 1-22 Special theory of relativity 1-13 Speed of light, measurement of 1-9 Speed of light wave 32-17 Spiraling electron in bubble chamber 28-27 The betatron 30-16 The early universe 34-29 Thought experiment on expanding magnetic field 32-11 Time dilation 1-22 Time-energy form of the uncertainty principle 40-19, 40-23 Zero rest mass int-12 Zero rest mass particles 6-11 Relativity, general int-15, 8-29

Renormalization int-14, int-17 Repeated wave forms in Fourier analysis 16-11 Resistors In parallel 27-12 In series 27-11 Introduction to 27-6 LR circuit 31-8 Ohm's law 27-7 Resonance Electron spin Classical picture of 39-14 Experiment 39-9 Introduction to 39-5 X. See Experiments II: -12- Electron spin resonance Introduction to 14-24 Phenomena 14-26 Tacoma Narrows bridge 14-24 Vortex street 14-25 Transients 14-27 Rest energy of proton and electron in eV 26-12 Rest mass int-11 And kinetic energy 10-5 Einstein formula 6-10 **Restoring forces** Linear 14-7 Non linear 14-19 Retrograde motion of Mars 8-24 Reversible engines As thought experiment 18-13 Carnot cycle 18-17 Efficiency of 18-18 Rifle and Bullet, recoil 7-7 Right handed coordinate system 2-18 **Right-hand rule** For cross products 12-10 For Faraday's law 30-15 For magnetic field of a current 28-13 For magnetic field of a solenoid 29-14 For surfaces 29-16 Mirror images of Optics-6 Rods, nerve fibers in eye Optics-31 Rope, working with 13-10 Rotational motion Angular acceleration 12-3 Angular analogy 12-3 Angular velocity 12-2 Bicycle wheel as a collection of masses 12-5 Chapter on 12-1 Radian measure 12-2 Rolling down inclined plane 12-25 Rotational kinetic energy 12-22 Proof of theorem 12-26 Translation and rotation 12-24 Rubber, elasticity of 17-35 Rutherford and the nucleus 35-1 Rydberg constant, in Bohr theory 35-9

# S

Sailboats, Bernoulli's equation 23-14 Salt Dissolving 17-6 Ionic bonding 38-15 Satellite motion 8-8 Calculational loop for 8-19 Classical hydrogen atom 35-2 Compare with projectile, Newton's sketch 8-10 Computer lab 8-23 Computer prediction of 8-16 Conservation of angular momentum 8-32 Conservation of energy 8-35 Earth tides 8-12 Gravitational potential energy 10-22 Kepler's laws 8-24 Kepler's first law 8-26 Kepler's second law 8-27 Kepler's third law 8-28 Modified gravity 8-29 Moon 8-8 Orbit, circular 10-27 Orbit, elliptical 10-27 Orbit, hyperbolic 10-27 Orbit, parabolic 10-27 Planetary units for 8-14 Program for (Orbit 1) 8-21 Total energy 10-26 X. See Experiments I: - 5- Computer analysis of satellite motion Scalar dot product 2-12 Definition of work 10-13 Scanning tunneling microscope Optics-51 Surface (111 plane) of a silicon Optics-51 Scattering of waves By graphite crystal, electron waves 36-8 By myoglobin molecule 36-5 By small object 36-2 By thin crystals 36-6 Chapter on 36-1 Davisson-Germer experiment 35-12 Reflection of light 36-3 Two slit thought experiment 40-10 X ray diffraction 36-4 Schmidt, Maarten, guasars 34-19 Schrödinger wave equation Felix Block story on 37-1 Introduction to wave motion 15-1 Particle-wave nature of matter int-10 Solution for hydrogen atom 38-2 Standing waves in fuzzy walled box 38-1 Schrödinger, Erwin int-10, 37-1 Schwinger, J., guantum electrodynamics int-14 Search coil For magnetic field mapping experiment 30-24 Inside Helmholtz coils 39-12 Second law, Newton's. See Newton's laws: Second law

Second law of thermodynamics. See also Carnot cycle; Thermal energy Applications of 18-17 Chapter on 18-1 Statement of 18-4 Time reversed movie 18-1 Second, unit of time, definition of int-2 Secondary mirror, in telescope Optics-42 Semi major axis, Kepler's laws 8-28 Series expansions Cal 1-23 Binomial Cal 1-23 Exponential function e to the x Cal 1-28 Series, harmonic 16-3 Series wiring Capacitors 27-21 Resistors 27-11 Set Window, BASIC computer command 5-7 Short circuit 27-9 Short rod, electric force exerted by 24-9 Silicon, surface (111 plane) of Optics-51 Simple electric circuit 27-8 Simple pendulum Simple and conical pendulums 14-17 Theory of 14-15 Simultaneity, lack of 1-32 Sine function Amplitude of Cal 1-37 Definition of Cal 1-35, Cal 1-36 Derivative of, derivation Cal 1-38 Sine waves AC voltage generator 30-22 Amplitude of 14-10 As solution of differential equation 14-9 Definition of 14-3 Derivative of 14-8 Formation of pulse from 40-27 Fourier analysis lecture 16-28 Fourier analysis of 16-7 Harmonic series 16-3 Normal modes 16-4 Phase of 14-6 Pulse in air 15-3 Sinusoidal waves motion 15-12 Standing waves on a guitar string 15-20 Traveling wave 15-16 Traveling waves add to standing wave 15-20 Single slit diffraction 33-26 Analysis of pattern 33-27 Application to uncertainty principle 40-16 Huygens principle 33-4 Recording patterns 33-28 Slit pattern, Fourier transform of 16-33 Slope of a curve As derivative Cal 1-30 Formula for Cal 1-30 Negative slope Cal 1-31

Small angle approximation Simple pendulum 14-16 Snell's law Optics-19 Small oscillations For non linear restoring forces 14-19 Molecular forces act like springs 14-20 Of simple pendulum 14-16 Smallest scale of distance, physics at int-25 Snell's law Applied to spherical surfaces Optics-19 Derivation of Optics-12 For small angles Optics-19 Introduction to Optics-2, Optics-11 Sodium to argon, periodic table 38-13 Solar neutrinos 6-13. See also Neutrinos Solenoid Ampere's law applied to 29-15 Magnetic field of 28-17, 29-14 Right hand rule for 29-14 Toroidal, magnetic field of 29-17 Solving problems Gauss' law problems 24-26 How to go about it 24-29 Projectile motion problems 4-16 Sombrero galaxy int-2 Sound Focusing by an ellipse 8-26, Optics-3 Fourier analysis, and normal modes 16-1 Fourier analysis of violin notes 16-18 Intensity Bells and decibels 16-24 Definition of 16-24 Speaker curves 16-27 Percussion instruments 16-22 Sound meters 16-26 Sound produced by guitar string 15-22 Stringed instruments 16-18 The human ear 16-16 Wind instruments 16-20 X. See Experiments I: -12- Fourier analysis of sound waves Sound waves, speed of Calculation of 15-8 Formula for 15-9 In various materials 15-9 Space And time int-2. 1-1 Quantum fluctuations in 40-25 The Lorentz contraction 1-24 Travel 1-22 Space physics 28-31. See also Experiments II: -5a- Magnetic focusing and space physics Space telescope Hubble Optics-44 Infrared (IRAS) Optics-47 Spacial frequency k 15-14 Speaker curves 16-27

Specific heat Cp and Cv 18-6 Definition of 17-26 Failure of Newtonian mechanics 17-31 Gamma = Cp/Cv 18-7 Spectral lines Atomic spectra 33-16 Bohr's explanation of int-9 Hydrogen Bohr theory of 35-4 Colors of int-7 Experiment to measure 33-17 The Balmer Series 33-19 Introduction to int-7 Spectrometer, mass 28-28 Spectrum Electromagnetic Photon energies 34-11 Visible spectrum 33-15 Wavelengths of 32-20 Hydrogen Balmer series 33-19 Bohr theory of 35-4 Experiment to measure 33-17 Lyman series, ultraviolet 35-6 Paschen series, infrared 35-6 Hydrogen star 33-19, 35-4 Spectrum. Electromagnetic Photon energies 34-11 Speed and mass increase int-11 Speed detector, air cart 30-5. See also Experiments II: - 6- Faraday's law air cart speed detector Speed limit, nature's int-12, 6-11 Speed of an electromagnetic pulse 32-14 Speed of light Absolute speed limit int-2, int-12 Calculation of speed of light Analysis of path 1 32-14 Analysis of path 2 32-16 Using Maxwell's Equations 32-14 Dimensional analysis 15-9 Experiment to measure 1-9, 31-15 In a medium Optics-8 Same to all observers 1-12 Speed of sound Formula for and values of 15-9 Theory of 15-8 Speed of wave pulses Dimensional analysis 15-6 On rope 15-4 Sphere Area of 23-6 Electric field inside of 26-4 Spherical aberration Optics-21 In Hubble telescope mirror Optics-22 Spherical lens surface Optics-19 Formula for focal length Optics-20

Spherical mass, gravitational field of 24-24 Spin Allowed projections 39-3 Chapter on 39-1 Concept of 39-3 Dirac equation 39-3 Electron, introduction to 38-1 Electron, periodic table 38-9 Electron Spin Resonance Experiment 39-5, 39-9 Gyroscope like 39-15 Interaction with magnetic field 39-4 Magnetic energy 39-1 Semi classical formula 39-14 Uhlenbeck and Gaudsmit 39-1 X. See Experiments II: -12- Electron spin resonance Spin flip energy, Dirac equation 39-15 Spring Constant 9-3 Forces 9-3 Mass on a spring Computer analysis of 14-30 Differential equation for 14-8 Theory 14-7 Oscillating cart 14-5 Spring model of molecular force 17-13 Spring pendulum Ball spring program 9-20 Computer analysis of 9-8 F = ma 9-7 Introduction to 9-4 Spring potential energy int-11 Square of amplitude, intensity 16-33 Square wave Fourier analysis of 16-9, 16-28 Stability of matter 19-14 Standard model of basic interactions int-25 Standing waves Allowed standing waves in hydrogen 37-1 De Broglie waves Movie 35-11 Formulas for 15-20 Hydrogen, L= 0 patterns 38-4 Introduction to 15-18 Light waves in laser 37-2 Made from traveling waves 15-19 On a guitar string 15-20 Frequency of 15-21 On drums 16-22 On violin backplate 16-23 Particle wave nature int-13 Patterns in hydrogen 38-3 Photons in laser 37-3 The L= 0 patterns in hydrogen 38-4 Two dimensional 37-8 Electrons on copper crystal 37-9 On drumhead 37-8

Star Binary int-2 Black hole 20-19 Black dwarf int-19 Black hole int-20, 20-18 Blackbody spectrum of 34-3 Hydrogen spectrum of 33-19, 35-4 Neutron int-20, 20-17 Red supergiant 20-15 Stellar evolution 17-17 White dwarf 20-15 Statamps, CGS units 24-2 Static charges Electric field of 30-16 Line integral for 30-2 Stationary source, moving observer, Doppler effect 33-21 Statvolts, CGS units of charge 24-2 Steady state model of the universe 34-25 Stellar evolution General discussion int-19 Role of neutrinos 20-15 Role of the four basic interactions 20-13 Role of thermal energy 17-17 Step-By-Step Calculations 5-1 Stomach, medical image Optics-15 Strain, definition of 15-8 Strange quark int-24 Streamlines And electric field lines 24-14, 24-17 Around airplane wing 23-13 Around sailboat sail 23-14 Bernoulli's equation 23-11 Bounding flux tubes 23-8 Definition of 23-4 Hele-Shaw cell 23-4 In blood flow experiment 23-23 In superfluid helium venturi meter 23-17 In venturi meter 23-15 Strength of the electric interaction 19-8 And magnetic forces 28-6 In comparison to gravity int-6 In comparison to nuclear force int-18, 20-2 Two garden peas 28-2 Stress, definition of 15-8 String forces Atwoods machine 9-16 Conical pendulum 9-18 Introduction to 9-15 Solving pulley problems 9-16 Working with rope 13-10 String theory int-16 Stringed instruments 16-18 Violin, acoustic vs electric 16-19

Strobe photographs Analyzing 3-8, 3-11 And the uncertainty principle 4-2, Cal 1-4 Defining the acceleration vector 3-15 Defining the velocity vector 3-11 Taking 3-7 Structure, nuclear 20-7 Styrofoam projectile 5-28 SU3 symmetry 19-15 Gell-Mann and Neuman int-24 Subtraction of vectors 2-7 Summation Becoming an integral Cal 1-10 Of velocity vectors Cal 1-10 Sun Age of 34-23 And neutron stars 20-17 Energy source int-18 Halos around Optics-18 Kepler's laws 8-24 Neutrinos from 6-13, 11-21 Stellar evolution 17-17, 20-13 Sun dogs Optics-18 Superconducting magnets 31-30 Superfluids Quantized vortices in 23-22 Superfluid helium venturi meter 23-15 Supernova, 1987 In stellar evolution 20-14 Neutrino astronomy 6-14 Neutrinos from 20-16 Photograph of int-19, 6-14 Superposition of waves Circular waves 33-2 Principle of 15-11 Two slit experiment 40-2 Surface (111 plane) of silicon Optics-51 Surface charges 26-2 Surface integral. See also Integral Definition of 29-2 For magnetic fields 32-3 Gauss' law 29-3 In Maxwell's equations 32-8 Two kinds of vector fields 30-19 Surface tension 17-14 Symmetry SU3 19-15 Gell-Mann and Neuman int-24 Symmetry of Maxwell's equations 32-9 Synchrotron 28-22 Systems of particles, chapter on 11-1

#### T

Table of elements 19-5 Tacoma narrows bridge 14-24 Tangential distance, velocity and acceleration 12-4 Tau particle, lepton family int-22 Tau type neutrino int-22 Taylor, Joe, binary neutron stars int-15 Telescope, parabolic reflector Optics-4 Telescopes int-1 Infrared IRAS satellite Optics-47 Mt. Hopkins, 2Mass Optics-46 Radio Optics-48 Arecibo Optics-48 Arecibo, binary neutron stars int-15 Holmdel, three degree radiation int-30 Radio galaxy image Optics-48 Radio images of variable star Optics-49 Very Large Array Optics-48 Very Long Baseline Array Optics-49 Reflecting Optics-42 Cassegrain design Optics-42 Diffraction limit Optics-45 Hubbel space telescope Optics-44 Image plane Optics-5 Keck, world's largest optical Optics-45 Mt. Hopkins Optics-43 Mt. Palomar Optics-43 Newton's Optics-42 Secondary mirror Optics-42 William Hershel's Optics-43 Refracting Optics-40 Galileo's Optics-41 Yerkes Optics-41 Television waves Photon energies 34-11 Wavelength of 32-20 Temperature Absolute zero 17-9 And zero point energy 37-8 Boltzman's constant 17-11 Heated hydrogen int-9 Ideal gas thermometer 17-20 Introduction to 17-9 Temperature scales 17-10 Tesla and Gauss, magnetic field dimensions 28-16 Test charge, unit size 24-11 Text file for FFT data 16-32 Thermal efficiency of Carnot cycle. See Carnot cycle, efficiency of Thermal energy Dollar value of 18-1 In a bottle of hydrogen int-8 Time reversed movie of dive 18-1 Thermal equilibrium And temperature 17-9 Introduction to 17-8 Of the universe 34-28

#### Index-36

Thermal expansion Adiabatic expansion 18-9 Derivation of formula 18-26 Formula for 18-10 In stellar evolution 17-17, 20-13 Isothermal expansion 18-8 Derivation of formula for 18-26 Molecular theory of 17-33 Of gas in balloon 17-16 Work done by an expanding gas 18-5 Thermal motion Boltzman's constant 17-11 Brownian motion 17-6 Movie 17-7 Thermal photons. See Three degree radiation In blackbody radiation 34-22 Thermometer Ideal gas 17-20 Mercury or alcohol 17-9 Temperature scales 17-10 Thin lenses Optics-23 Thought experiments Carnot cycle 18-4 Causality 1-36 Lack of simultaneity 1-32 Light pulse clock 1-13 Lorentz contraction and space travel 1-22 Magnetic force and Faraday's law 30-3 No width contraction 1-28 Origin of magnetic forces 28-8 Two slit experiment and the uncertainty principle 40-9 Three degree cosmic radiation int-30, 34-27 Penzias and Wilson 34-27 Tides, two a day 8-12 Time Age of sun 34-23 And the speed of light int-2 Behavior of 1-1 Dilation 1-22 Dilation formula 1-16 Direction of Entropy 18-26 Neutral K meson 18-26 In early universe int-27, 34-29 Measuring short times using uncertainty principle 40-22 Moving clocks 1-13 Muon lifetime movie 1-21 On light pulse clock 1-14 On other clocks 1-18 On real clocks 1-20 Steady state model of universe 34-25 Time constant For RC circuit 27-24 Measuring from a graph 27-25, Cal 1-34

Time, direction of And strobe photographs 3-27 Time reversal Dive movie 18-1 Water droplets 18-2 Time step and initial conditions 5-14 Time-energy form of the uncertainty principle 40-19 Top quark Mass of int-24 Quark family int-24 Toroid Inductor 31-6 In LC experiment 31-11 In resonance 31-14 Speed of light measuerment 1-9, 31-15 Magnetic field of 29-17 Torque As angular force 12-15 In torsion pendulum 14-12 On a current loop 31-20 Torsion pendulum 14-12. See also Cavendish experiment Total energy Classical hydrogen atom 35-3 Escape velocity 10-28 Satellite motion 8-36 Town water supply, hydrodynamic voltage 23-18 Transients 14-27 Translation and rotation 12-24 Transmitted wave and lenses 36-3 Traveling waves, formula for 15-16 Tritium, a hydrogen isotope int-17, 19-7 True BASIC. See BASIC program Tubes of flux 24-17 Two dimensional standing waves 37-8 Two kinds of vector fields 30-18 Two lenses together Optics-29 Two slit experiment And the uncertainty principle 40-9 Measurement limitations 40-9 One particle at a time 40-3 One slit experiment 40-2 Particle point of view 40-3 Particle/wave nature 40-2 Using electrons 40-3 Two-slit interference patterns 33-6 A closer look at 33-26 First maxima 33-8 For light 33-10 Tycho Brahe's apparatus 8-25

### U

Uhlenbeck and Gaudsmit, electron spin 39-1 Ultraviolet light *Electromagnetic spectrum int-7, 32-20, 32-22 Photon energies 34-11 Wavelength of 32-20* 

Uncertainty principle Cal 1-3  $\Delta x \Delta p >= h \quad 40-15$ And definition of velocity 4-1, Cal 1-3 And strobe photographs 4-2, Cal 1-4 Applied to projectile motion 4-2, Cal 1-4 Elementary particles, short lived 40-23 Energy conservation 40-24 Fourier transform 40-20 Introduction to 40-14 Particle/wave nature int-10 Position-momentum form 40-15 Single slit experiment 40-16 Time-energy form 40-19 Used as clock 40-22 X. See Experiments II: -13- Fourier analysis & the uncertainty principle Uniform circular motion. See Circular motion Uniform expansion of universe int-3, 33-24 Uniform magnetic fields 28-16 Unit of angular momentum int-9 Electron spin 39-3 In Bohr theory 35-9 Unit test charge 24-11 Unit vectors 8-18 Units Atomic units 19-22 CGS centimeter, gram, second Back cover-1 Coulomb's law 24-2 Statamp, statvolt, ergs per second 24-2 Checking MKS calculations 24-3 MKS Ampere, volt, watt 24-2 Coulomb's law 24-2 Meter, kilogram, second Front cover-2 Planetary units 8-14 Practical System of units (MKS) 10-31 Universe Age of int-3 As a laboratory int-1 Becomes transparent, decoupling 34-31 Big bang model int-4, 33-25, 34-26 Continuous creation theory int-4, 34-25 Decoupling (700,000 years) 34-31 Early. See Universe, early Evolution of, Doppler effect 34-21 Excess of matter over antimatter int-27, 34-29 Expanding int-3 Quasars and gravitational lensing 34-19 Red shift 33-24 Helium abundance in 34-26 Models of 34-23 Steady state model of 34-25 The First Three Minutes 34-32 Thermal equilibrium of 34-28 Three degree cosmic radiation int-30, 34-27 Visible int-3, 34-32

Universe, early int-27, 34-29 10 to the 10 degrees int-27 10 to the 13 degrees int-27 10 to the 14 degrees int-27 13.8 seconds int-28 24% neutrons int-28 38% neutrons int-27 At .7 million years, decoupling int-29 At various short times 34-30 Big bang model 34-26 Books on Coming of Age in the Milky Way 34-32 The First Three Minutes 34-32 Decoupling (700,000 years) 34-31 Deuterium bottleneck int-28 Excess of Matter over Antimatter 34-29 Frame #2 (.11 seconds) 34-30 Frame #3 (1.09 seconds) 34-30 Frame #4 (13.82 seconds) 34-30 Frame #5 (3 minutes and 2 seconds) 34-30 Helium abundance 34-26 Helium created int-28 Hot int-4 Matter particles survive int-27, 34-29 Neutrinos escape at one second int-28 Overview int-27. 34-29 Positrons annihilated int-28 Thermal equilibrium of 34-28 Thermal photons int-29, 34-28 Three degree cosmic radiation int-30, 34-27, 34-32 Transparent universe int-29 Why it is hot 34-29 Up quark int-24 Uranium Binding energy per nucleon 20-10 Nuclear fission 20-4 Nuclear structure int-17, 20-2

#### V

Van Allen radiation belts 28-32 Van de Graaff generator 26-6 Variable names, computer 5-6 Variables *Evaluated over interval Cal 1-10*  Vector Addition 2-3 Addition by components 2-9 Angular momentum 7-14, 12-7 Angular velocity 12-7 Area 24-22 Components 2-8 Cross Product 12-9 Definition of acceleration Cal 1-7 Component equations Cal 1-8 Definition of velocity Cal 1-6 Component equations Cal 1-8 Dot product 10-13 Equations Components with derivatives 4-6, Cal 1-7 Constant acceleration formulas 4-11 Exercise on 2-7, 2-8 In component form 2-10 Magnitude of 2-6 Measuring length of 3-9 Multiplication 2-11 Multiplication, cross product 2-15, 12-9 Formula for 2-17 Magnitude of 2-17 Right hand rule 12-10 Multiplication, scalar or dot product 2-12, 10-13 Interpretation of 2-14 Velocity from coordinate vector 3-13 Vector fields 23-3 Electric field 24-10 Magnetic field 28-10 Two kinds of 30-18 Velocity field 23-2 Vectors 2-2 Arithmetic of 2-3 Associative law 2-4 Commutative law 2-4 Coordinate 3-11 Displacement 2-2 from Strobe Photos 3-5 Graphical addition and subtraction 3-10 Multiplication by number 2-5 Negative of 2-5 Subtraction of 2-5 Unit 8-18 Velocity 3-11 And the uncertainty principle 4-2, Cal 1-4 Angular 12-2, 14-5 Angular analogy 12-3 Calculus definition of 4-3, Cal 1-5 Component equations Cal 1-8 Curve, area under Cal 1-12 Definite integral of Cal 1-11 Instantaneous 3-24 From strobe photograph 3-26 Integral of Cal 1-10 Of escape 10-28 Tangential 12-4 Using strobe photos 3-11

Velocity detector Air cart 30-5 Magnetic flux 30-25 Velocity field 23-2 Flux of 23-8, 24-15 Of a line source 23-7 Of a point source 23-6 Velocity vector from coordinate vector 3-13 Venturi meter Bernoulli's equation 23-15 With superfluid helium 23-17 Very Large Array, radio telescopes Optics-48 Very Long Baseline Array, radio telescopes Optics-49 Violets, odor of 17-6 Violin Acoustic vs electric 16-19 Back, standing waves on 16-23 Viscous effects in fluid flow 23-19 Visible light int-7, 32-20 Photon energies 34-11 Spectrum of 33-15 Wavelength of 32-20 Visible universe int-3, 34-32 Volt Electron 26-12 MKS units 24-2 Voltage Air cart speed detector 30-5 Divider, circuit for 27-13 Electric 25-6 Fluid analogy Hydrodynamic voltage 23-17 Resistance 27-7 Town water supply 23-18 Induced 31-4 Induced in a moving loop 30-4 Voltage and current 27-6 Resistors 27-6 Ohm's law 27-7 Voltage source, constant 27-15 Voltage transformer and magnetic flux 30-26 Volume of mole of gas 17-25 Vortex Bathtub 23-2, 23-20 Hurricane 23-20 Tornado 23-20 Vortex street, Karman 14-25 Vortices 23-20 Quantized, in superfluids 23-22
## W

W and Z mesons, electroweak interaction int-26 Water Evaporating 17-5 Molecule 17-2 Water droplets, time reversal 18-2 Watt. MKS units 24-2 Wave Circular water waves 33-2 Cosine waves 16-28 De Broglie, standing wave movie 35-11 Diffraction pattern 33-5 Electromagnetic waves 32-18 Probability wave for photons 40-7 Electron waves, de Broglie picture 35-11 Electron waves, in hydrogen 38-1 Forms, repeated, in Fourier analysis 16-11 Fourier analysis Of a sine wave 16-7 Of a square wave 16-9 Huygens' principle, introduction to 33-4 Interference patterns 33-3 Light waves Chapter on 33-1 Polarization of 32-23 Patterns, superposition of 33-2 Photon wave 40-6 Probability wave Intensity of 40-22 Reflection and Fluorescence 40-8 Scattering, measurement limitation 40-10 Single slit experiment, uncertainty principle 40-16 Sinusoidal waves 15-12 *Time dependent* 14-3 Speed of waves Dimensional analysis 15-6 Project suggestion 15-8 Standing waves Allowed 37-1 Formulas for 15-20 Frequency of 15-21 Hydrogen L= 0 Patterns 38-4 Introduction to 15-18 On a guitar string 15-20 Two dimensional 37-8 Transmitted waves 36-3 Traveling waves, formula for 15-16 Wave pulses 15-3 Speed of 15-4 Wave equation int-7, 15-1 Dirac's 15-2 Bohr magneton 39-5 Spin 39-3 For light int-7 Maxwell's 15-1 Nonrelativistic int-12, 34-16 Relativistic int-12, 15-2 Schrödinger's 15-1 Discovery of 37-1

Wave motion Amplitude and phase 15-17 Linear and nonlinear 15-10 One dimensional 15-1 Principle of superposition 15-11 Wave nature of light Young, Thomas 34-1, Optics-1 Wave patterns Hydrogen Intensity at the Origin 38-5 Schrödinger's equation 38-2 Standing waves 38-3 Wave/particle nature. See also Particle-wave nature Born Interpretation 40-6 Probability interpretation 40-6 Two slit experiment 40-2 Wavelength De Broglie 35-11 Electron 36-9 Fourier analysis 16-28 Laser beam 16-33 Period, and frequency 15-13 Weak interaction int-14, int-20, 19-2 Creation of neutrinos 20-6 Electroweak theory int-26 Four basic interactions 19-2 Neutron decay 20-6 Range of int-21 Strength of int-21 Weighing the earth 8-8 Weight 8-11 Lifting 13-11 Weinberg, S., book "The first Three Minutes" 34-32 Wein's displacement law for blackbody radiation 34-2 White dwarf star 20-15 White light (Fourier analysis) 16-28 William Hershel's telescope Optics-43 Wilson, Robert, 3 degree radiation int-29, 34-27 Wind instruments 16-20 Work int-11 And potential energy 10-14 Bernoulli's equation 23-10 Calculation of in adiabatic expansion 18-26 Calculation of in isothermal expansion 18-26 Definition of 10-12 Done by an expanding gas 18-5 Integral formula for 10-14, 10-15 Non-constant forces 10-14 Vector dot product 10-13 Work energy theorem 10-18 World's largest optical telescope Optics-45

## X

x-Cal 1 Exercise 1 Cal 1-14 Exercise 2 Cal 1-15 Exercise 3 Cal 1-17 Exercise 4 Cal 1-22 Exercise 5 Cal 1-24 Exercise 6 Cal 1-29 Exercise 7 Cal 1-29 Exercise 8 Cal 1-31 Exercise 9 Cal 1-33 Exercise 10 Cal 1-36 Exercise 11 Cal 1-39 Exercise 12 Cal 1-39 Exercise 13 Cal 1-39 X-Ch 1 Exercise 1 1-3 Exercise 2 1-4 Exercise 3 1-11 Exercise 4 1-16 Exercise 5 1-31 Exercise 6 1-31 Exercise 7 1-31 X-Ch 2 Exercise 1 2-7 Exercise 2 2-7 Exercise 3 2-7 Exercise 4 2-7 Exercise 5 2-8 Exercise 6 2-10 Exercise 7 2-12 Exercise 8 2-13 Exercise 9 2-15 Exercise 10 2-16 Exercise 11 2-17 Exercise 12 2-17 Exercise 13 2-18 X-Ch 3 Exercise 1 3-10 Exercise 2 3-10 Exercise 3 3-10 Exercise 4 3-12 Exercise 5 3-17 Exercise 6 3-18 Exercise 7 3-22 Exercise 8 3-27 Exercise 9 3-27 Exercise 10 3-27 X-Ch 4 Exercises 1-7 4-19 X-Ch 5 Exercise 1 5-3 Exercise 2 A running program 5-8 Exercise 3 Plotting a circular line 5-8 Exercise 4 Labels and axes 5-9 Exercise 5a Numerical output 5-9 Exercise 5b 5-10 Exercise 6 Plotting crosses 5-11

Exercise 7 5-20 Exercise 8 Changing the time step 5-20 Exercise 9 Numerical Output 5-20 Exercise 10 Attempt to reduce output 5-20 Exercise 11 Reducing numerical output 5-20 Exercise 12 Plotting crosses 5-21 Exercise 13 Graphical analysis 5-25 Exercise 14 Computer prediction 5-26 Exercise 15 Viscous fluid 5-26 Exercise 16 Nonlinear air resistance (optional) 5-27 Exercise 17 Fan Added 5-27 X-Ch 6 Exercise 1 6-4 Exercise 2 Decay of Plutonium 246 6-8 Exercise 3 Protactinium 236 decay. 6-9 Exercise 4 Increase in Electron Mass. 6-9 Exercise 5 A Thought Experiment. 6-9 Exercise 6 6-10 Exercise 7 6-10 Exercise 8 6-10 X-Ch 7 Exercise 1 7-6 Exercise 2 7-8 Exercise 3 Frictionless Ice 7-8 Exercise 4 Bullet and Block 7-8 Exercise 5 Two Skaters Throwing Ball 7-8 Exercise 6 Rocket 7-8 Exercise 7 7-10 Exercise 8 7-10 Exercise 9 7-10 Exercise 10 7-13 Exercise 11 7-14 Exercise 12 7-16 Exercise 13 7-16 Exercise 14 7-18 X-Ch 8 Exercise 1 8-5 Exercise 2 8-8 Exercise 3 8-9 Exercise 4 8-11 Exercise 5 8-12 Exercise 6 8-15 Exercise 7 8-15 Exercise 8 8-16 Exercise 9 8-16 Exercise 10 8-22 Exercise 11 8-23 Exercise 12 8-23 Exercise 13 8-27 Exercise 14 8-27 Exercise 15 8-28 Exercise 16 8-28 Exercise 17 8-28 Exercise 18 8-31 Exercise 19 8-34 Exercise 20 8-34 Exercise 21 8-36 Exercise 22 8-36 Exercise 23 8-37

X-Ch 9 Exercise 1 9-7 Exercise 2 9-11 Exercise 3 9-17 Exercise 4 9-17 Exercise 5 Conical Pendulum 9-19 Exercise 6 9-19 X-Ch10 Exercise 1 10-4 Exercise 2 10-6 Exercise 3 10-8 Exercise 4 10-9 Exercise 5 10-9 Exercise 6 10-9 Exercise 7 10-11 Exercise 8 10-12 Exercise 9 10-13 Exercise 10 10-17 Exercise 11 10-17 Exercise 12 10-28 Exercise 13 10-28 Exercise 14 10-28 Exercise 15 10-31 X-Ch11 Exercise 1 11-4 Exercise 2 11-4 Exercise 3 11-4 Exercise 4 11-6 Exercise 5 11-8 Exercise 6 11-12 Exercise 7 11-13 Exercise 8 11-13 Exercise 9 11-13 Exercise 10 11-14 Exercise 11 11-14 Exercise 12 11-15 Exercise 13 11-16 X-Ch12 Exercise 1 12-2 Exercise 2 12-3 Exercise 3 12-5 Exercise 4 12-9 Exercise 5 12-10 Exercise 6 12-11 Exercise 7 12-13 Exercise 8 12-15 Exercise 9 12-21 Exercise A1 12-24 Exercise A2 12-26 Exercise A3 potential lab experiment 12-26 X-Ch13 Exercise 1 13-3 Exercise 2 13-3 Exercise 3 13-4 Exercise 4 13-6 Exercise 5 13-8 Exercise 6 Ladder problem 13-8 Exercise 7 13-9

Exercise 8 Working with rope 13-10 Exercise 9 13-11 Exercise 10 13-12 Exercise 11 13-12 X-Ch14 Exercise 1 14-4 Exercise 2 14-5 Exercise 3 14-5 Exercise 4 14-7 Exercise 5 14-8 Exercise 6 14-9 Exercise 7 14-10 Exercise 8 14-10 Exercise 9 14-10 Exercise 10 14-11 Exercise 11 14-11 Exercise 12 14-14 Exercise 13 14-14 Exercise 14 14-15 Exercise 15 14-16 Exercise 16 14-16 Exercise 17 14-17 Exercise 18 14-18 Exercise 19 Physical pendulum 14-18 Exercise 20 Damped harmonic motion 14-23 Exercise 21 14-27 Exercise 22 14-33 Exercise 23 14-34 X-Ch15 Exercise 1 15-4 Exercise 2 15-9 Exercise 3 15-9 Exercise 4 15-9 Exercise 5 15-14 Exercise 6 15-15 Exercise 7 15-16 Exercise 8 15-22 Exercise 9 15-22 Exercise 10 15-22 X-Ch16 Exercise 1 16-21 Exercise 2 16-27 Exercise 3 16-27 X-Ch17 Exercise 1 17-8 Exercise 2 17-11 Exercise 3 17-11 Exercise 4 17-11 Exercise 5 17-14 Exercise 6 17-24 Exercise 7 17-25 Exercise 8 17-26

X-Ch18 Exercise 1 18-5 Exercise 2 18-7 Exercise 3 18-10 Exercise 4 18-13 Exercise 5 18-13 Exercise 6 18-14 Exercise 7 18-16 Exercise 8 18-19 Exercise 9 18-20 Exercise 10 18-20 X-Ch19 Exercise 1 19-9 Exercise 2 19-11 Exercise 3 19-11 Exercise 4 19-12 Exercise 5 19-12 X-Ch20 Exercise 1 20-9 X-Ch23 Exercise 1 23-12 Exercise 2 23-12 Exercise 3 23-15 Exercise 4 23-15 Exercise 5 23-22 Exercise 6 23-22 X-Ch24 Exercise 1 24-5 Exercise 2 24-5 Exercise 3 24-5 Exercise 4 24-5 Exercise 5 24-8 Exercise 6 24-9 Exercise 7 24-12 Exercise 8 24-25 Exercise 9 24-27 Exercise 10 24-27 Exercise 11 24-27 Exercise 12 24-27 Exercise 13 24-28 Exercise 14 24-28 X-Ch25 Exercise 1 25-5 Exercise 2 25-9 Exercise 3 25-12 Exercise 4 25-12 X-Ch26 Exercise 1 26-5 Exercise 2 26-5 Exercise 3 26-5 Exercise 4 26-5 Exercise 5 26-11 Exercise 6 26-13 Exercise 7 26-13 Exercise 8 26-16 Exercise 9 26-17 Exercise 10 Millikan oil drop experiment 26-17 X-Ch27 Exercise 1 27-10 Exercise 2 27-13 Exercise 3 The voltage divider 27-13 Exercise 4 - Electrolytic capacitor 27-17 Exercise 5 27-19 Exercise 6 27-21 Exercise 7 27-24 Exercise 8 27-26 Exercise 9 27-28 Exercise 10 27-30 Exercise 11 27-32 X-Ch28 Exercise 1 28-3 Exercise 2 28-9 Exercise 3 28-15 Exercise 4 28-21 Exercise 5 28-21 Exercise 6 28-25 Exercise 7 28-25 Exercise 8 28-25 Exercise 9 28-27 Exercise 10 28-28 X-Ch29 Exercise 1 29-12 Exercise 2 29-12 Exercise 3 29-12 Exercise 4 29-12 Exercise 5 29-13 Exercise 6 29-13 Exercise 7 29-16 Exercise 8 29-18 Exercise 9 29-18 X-Ch30 Exercise 1 30-10 Exercise 2 30-12 Exercise 3 30-18 Exercise 4 30-20 Exercise 5 30-22 Exercise 6 30-22 Exercise 7 30-25 Exercise 8 30-26 X-Ch31 Exercise 1 31-3 Exercise 2 31-9 Exercise 3 31-10 Exercise 4 31-11 Exercise 5 31-15 Exercise 6 31-15 Exercise 7 31-16 X-Ch32 Exercise 1 32-7 Exercise 2 32-8 Exercise 3 32-9 Exercise 4 32-17 Exercise 5 32-17 Exercise 6 32-32

X-Ch33 Exercise 1 33-4 Exercise 2 33-8 Exercise 3 33-10 Exercise 4 33-10 Exercise 5 33-13 Exercise 6 33-13 Exercise 7 33-16 Exercise 8 33-18 Exercise 9 33-19 Exercise 10 33-21 Exercise 11 33-23 Exercise 12 33-23 Exercise 13 33-23 Exercise 14 33-28 Exercise 15 33-30 X-Ch34 Exercise 1 34-3 Exercise 2 34-9 Exercise 3 34-9 Exercise 4 34-10 Exercise 5 34-10 Exercise 6 34-10 Exercise 7 34-10 Exercise 8 34-10 Exercise 9 34-10 Exercise 10 34-10 Exercise 11 34-17 Exercise 12 34-18 x-Ch35 Exercise 1 35-5 Exercise 2 35-5 Exercise 3 35-6 Exercise 4 35-6 Exercise 5 35-7 Exercise 6 35-9 Exercise 7 35-9 Exercise 8 35-9 Exercise 9 35-10 Exercise 10 35-10 x-Ch36 Exercise 1 36-3 Exercise 2 36-7 Exercise 3 36-7 Exercise 4 36-9 Exercise 5 36-12 Exercise 6 36-12 x-Ch37 Exercise 1 37-4 Exercise 2 37-6 Exercise 3 37-6 Exercise 4 37-7 x-Ch38 Exercise 1 38-4 Exercise 2 38-7

x-Ch39 Exercise 1 39-5 Exercise 2 39-5 Exercise 3 39-5 Exercise 4 39-7 Exercise 5 39-7 Exercise 6 39-12 x-Ch40 Exercise 1 40-6 Exercise 2 40-7 Exercise 3 40-18 X-Optics Exercise 1a Optics-10 Exercise 1b Optics-10 Exercise 2 Optics-12 Exercise 3 Optics-13 Exercise 4 Optics-17 Exercise 5 Optics-21 Exercise 6 Optics-21 Exercise 7 Optics-21 Exercise 8 Optics-23 Exercise 9 Optics-24 Exercise 10 Optics-27 Exercise 11 Optics-29 Exercise 12 Optics-30 Exercise 13 Optics-30 X-rays Diffraction 36-4 Diffraction pattern 36-5 Electromagnetic spectrum int-7, 32-20, 32-22 Photon energies 34-11, 36-4 Wavelength of 32-20

## Y

Yerkes telescope Optics-41 Young, Thomas *Wave nature of light 34-1, Optics-1* Young's modulus 15-8 Yukawa, H. int-22 Yukawa's theory int-22

## Z

Z and W mesons, electroweak interaction int-26 Z, astronomer's Z factor for Doppler effect 33-23 Zero, absolute 17-9, 17-21 Zero of potential energy 10-22 Zero point energy 37-7 *And temperature 37-8 Chapter on 37-1* Zero rest mass particles 6-11 Zoom lens Optics-18 Zweig, George, quarks 19-15

## Physical Constants in CGS Units (link to MKS units)

speed of light	$c = 3 \times 10^{10} \text{ cm/sec} = 1000 \text{ ft/}\mu\text{sec} = 1 \text{ ft/}nanosecond$
acceleration due to gravity	
at the surface of the earth	$g = 980 \text{ cm/sec}^2 = 32 \text{ ft/sec}^2$
gravitational constant	$G = 6.67 \times 10^{-8} \text{ cm}^3/(\text{gm sec}^2)$
charge on an electron	$e = 4.8 \times 10^{-10} esu$
Planck's constant	$h = 6.62 \times 10^{-27} \text{ erg sec} (\text{gm cm}^2/\text{sec})$
Planck constant $/2\pi$	$\hbar = 1.06 \times 10^{-27} \text{erg sec} (\text{gm cm}^2 / \text{sec})$
Bohr radius	$a_0 = .529 \times 10^{-8} cm$
rest mass of electron	$m_e = 0.911 \times 10^{-27} gm$
rest mass of proton	$M_{\rm p} = 1.67 \times 10^{-24} {\rm gm}$
rest energy of electron	$m_{e}^{r}c^{2} = 0.51 \text{ MeV} (\approx 1/2 \text{ MeV})$
rest energy of proton	$M_{p}c^{2} = 0.938 \text{ BeV} (\approx 1 \text{ BeV})$
proton radius	$r_{p} = 1.0 \times 10^{-13} cm$
Boltzmann's constant	$\dot{k} = 1.38 \times 10^{-16}$ ergs/kelvin
Avogadro's number	$N_0 = 6.02 \times 10^{23}$ molecules/mole

absolute zero =  $0^{\circ}K = -273^{\circ}C$ density of mercury =  $13.6 \text{ gm} / \text{cm}^3$ mass of earth =  $5.98 \times 10^{27} \text{gm}$ mass of the moon =  $7.35 \times 10^{25} \text{gm}$ mass of the sun =  $1.97 \times 10^{33} \text{gm}$ earth radius =  $6.38 \times 10^8 \text{cm} = 3960 \text{ mi}$ moon radius =  $1.74 \times 10^8 \text{cm} = 1080 \text{ mi}$ mean distance to moon =  $3.84 \times 10^{10} \text{cm}$ mean distance to sun =  $1.50 \times 10^{13} \text{cm}$ mean earth velocity in orbit about sun = 29.77 km / sec

### **Conversion Factors**

1 meter = 100 cm (100 cm/meter) 1 in. = 2.54 cm (2.54 cm/in.)1 mi = 5280 ft (5280 ft/mi) 1 km (kilometer) =  $10^5$  cm ( $10^5$  cm / km) 1 mi = 1.61 km =  $1.61 \times 10^5$  cm ( $1.61 \times 10^5$  cm/mi) 1 Å (angstrom) =  $10^{-8}$  cm ( $10^{-8}$  cm / Å)  $1 \text{ day} = 86,000 \text{ sec} (8.6 \times 10^4 \text{ sec} / \text{ day})$ 1 year =  $3.16 \times 10^7$  sec (3.16×10<sup>7</sup> sec/year) 1  $\mu$  sec (microsecond ) = 10<sup>-6</sup>sec (10<sup>-6</sup>sec /  $\mu$  sec ) 1 nanosecond = 10<sup>-9</sup>sec (10<sup>-9</sup>sec /nanosecond ) 1 mi/hr = 44.7 cm/sec60 mi/hr = 88 ft/sec  $1 \text{ kg} (\text{kilogram}) = 10^{3} \text{gm} (10^{3} \text{gm} / \text{kg})$ 1 coulomb =  $3 \times 10^9$  esu ( $3 \times 10^9$  esu/coulomb) 1 ampere =  $3 \times 10^9$  statamps ( $3 \times 10^9$  statamps/ampere) 1 statvolt = 300 volts (300 volts/statvolt) 1 joule =  $10^7$  ergs ( $10^7$  ergs / joule)  $1 \text{ W} (\text{watt}) = 10^7 \text{ ergs/ sec} (10^7 \text{ erg} / \text{W})$  $1 \text{ eV} = 1.6 \times 10^{-12} \text{ ergs}$  ( $1.6 \times 10^{-12} \text{ ergs/eV}$ )  $1 \text{ MeV} = 10^{6} \text{eV} (10^{6} \text{eV}/\text{MeV})$  $1 \text{ BeV} = 10^9 \text{eV}$  ( $10^9 \text{eV}$  /BeV)  $1 \mu$  (micron ) pressure = 1.33 dynes / cm<sup>2</sup> 1 cm Hg pressure =  $10^4 \mu$  $1 \text{ atm} = 76 \text{ cm Hg} = 1.01 \times 10^{6} \text{ dynes/cm}^{2}$ 



Moose Mountain Digital Press

# Calculus 2000

**E. R. Huggins** Dartmouth College



physics2000.com

Ce

## **MKS Units**

m = meters I	kg = kilograms	s = seconds			
N = newtons J	l = joules	C = coulombs			
T = tesla	F = farads	H = henrys	Powers	s of 10	
A = amperes l	K = kelvins	mol = mole	Derman	Ductor	Currench e l
speed of light	с	$3.00 \times 10^8 \mathrm{m}/\mathrm{s}$		Prelix	эутрог
gravitational cor	nstant G	$6.67 \times 10^{-11} \mathrm{N} \cdot \mathrm{m}^2  /  \mathrm{kg}^2$	10 10 <sup>9</sup>	tera	I C
permittivity con	stant $\varepsilon_0$	$8.85 \times 10^{-12} \mathrm{F/m}$	10	giga	G
	otont II	$1.26 \times 10^{-6} \text{H}/\text{m}$	10°	mega	Μ
permeability con	Istant $\mu_0$	1.20 × 10 H/III	$10^{3}$	kilo	k
elementary char	ge e	$1.60 \times 10^{-19}$ C	$10^{2}$	hecto	h
electron volt	eV	$1.60 \times 10^{-19} \text{J}$	$10^{-1}$	deci	d
electron rest mas	ss m <sub>e</sub>	$9.11 \times 10^{-31}$ kg	$10^{-2}$	centi	c
proton rest mass	m <sub>p</sub>	$1.67 \times 10^{-27} \text{kg}$	$10^{-3}$	milli	m
Planck constant	h	$6.63 \times 10^{-34} \text{J} \cdot \text{s}$	$10^{-6}$	micro	μ
Planck constant	/2π ħ	$1.06 \times 10^{-34} \text{J} \cdot \text{s}$	10 <sup>-9</sup>	nano	n
Bohr radius	r <sub>b</sub>	$5.29 \times 10^{-11} \text{m}$	$10^{-12}$	pico	р
Bohr magneton	$\mu_{\rm b}$	$9.27 \times 10^{-24} \text{J} / \text{T}$	$10^{-15}$	femto	f
Boltzmann cons	tant k	$1.38 \times 10^{-23} \text{J} / \text{K}$			
Avogadro const	ant N <sub>A</sub>	$6.02 \times 10^{23} \text{mol}^{-1}$			
universal gas co	nstant R	8.31 J/mol·K			

## Dimensions

Quantity	Unit		Equivalents	
Force	newton	Ν	J/m	kg•m/s $^2$
Energy	joule	J	N•m	$kg \cdot m^2/s^2$
Power	watt	W	J/s	kg•m <sup>2</sup> /s <sup>3</sup>
Pressure	pascal	Pa	N/m <sup>2</sup>	$kg/m \cdot s^2$
Frequency	hertz	Hz	cycle/s	s <sup>-1</sup>
Electric charge	coulomb	С		A•s
Electric potential	volt	V	J/C	kg•m <sup>2</sup> /A•s <sup>3</sup>
Electric resistance	ohm	Ω	V/A	kg•m <sup>2</sup> /A <sup>2</sup> •s <sup>3</sup>
Capacitance	farad	F	C/V	$A^2 \cdot s^4 / kg \cdot m^2$
Magnetic field	tesla	Т	N•s/C•m	$kg/A \cdot s^2$
Magnetic flux	weber	Wb	$T \cdot m^2$	$kg \cdot m^2 / A \cdot s^2$
Inductance	henry	Н	V•s/A	kg•m <sup>2</sup> /A <sup>2</sup> •s <sup>2</sup>

Copyright © Moose Mountain Digital Press New Hampshire 03750 All rights reserved

## Calculus 2000 A Physics-Based Calculus Text

When developing a physics curriculum, a major concern is the mathematical background of the student. The **Physics 2000** text was developed teaching premedical students who were supposed to have had one semester of calculus. Because many of the students had taken calculus several years previously, and had forgotten much of it, the physics text used strobe photographs and the computer to carefully introduce the calculus concepts such as velocity, acceleration, and the limiting process. By the time we got to electricity and magnetism in **Part 2** of **Physics 2000** we relied on the student being familiar with the basic steps of differentiation and integration.

For students who have forgotten much of their calculus course, or those who have not had calculus but wish to study the Physics 2000 text, we have written Chapter 1 of **Calculus 2000**. This chapter not only covers all the calculus needed for the **Physics 2000** text, but is also carefully integrated with it. The chapter is much shorter than the typical introductory calculus text because the basic calculus concepts are discussed in the physics text and the calculus chapter only has to deal with the formalism. After the introductory courses, the standard physics curriculum repeatedly goes over the same topics at successively higher mathematical levels. A typical example is the subject of electricity and magnetism which is taught using integral equations in the introductory course, using differential operators in an upper level undergraduate course, and then taught all over again in a graduate level course. In each of the courses it takes a while for the student to realize that this is just the same old subject dressed up in new math.

With Chapters 2 through 13 of the **Calculus 2000**, we introduce a different approach. We take the topics that we have already introduced in **Physics 2000**, and show how these topics can be handled in progressively more sophisticated mathematical ways. Once we have introduced the mathematical concepts of gradient, divergence and curl in the calculus text, we can turn the integral form of Maxwell's equation into a wave equation for electric and magnetic fields. With the introduction of the Laplacian and complex variables, we can study Schrödinger's equation and begin to solve for the hydrogen wave patterns discussed in Chapter 38 of the physics text. Beyond seeing the same topics in a more sophisticated way, the student finds that new insights can result from the advanced mathematical approach. Chapter 10 of the calculus text is a short chapter less than two pages. But it is one of the most significant chapters in the text. For there we see that **Maxwell's** equations for electric and magnetic fields require that electric charge be conserved. This intimate connection between a conservation law and field theory becomes clear when we have sufficiently powerful mathematical tools to handle the theory.

The physics text began its discussion of vector fields in Chapter 23, using the velocity field as its first example. We did that because it is much easier to visualize the familiar flow of water than the abstract concept of an electric field. We saw that the streamlines in fluid flow went over to electric field lines, Gauss's law in fluid theory simply represented the incompressibility of the fluid, and Bernoulli's equation provided an introduction to the concept of voltage and potential. However our discussion of electric and magnetic fields, particularly in this calculus text, go way beyond the simple fluid flow topics we introduced in the physics text. In the last two chapters of the calculus text, we turn the tables and apply to fluid theory the mathematical techniques we learned studying electricity and magnetism. In Chapter 12 we discuss the concept of vorticity which is the curl of the velocity field. The focus is to develop an intuitive understanding of the nature of vorticity and the role it plays in fluid flows, particularly vortices and vortex rings.

Chapter 13 is an introduction to fluid dynamics. The idea is to bring our discussion of the velocity field up to the same level as our treatment of electric and magnetic fields. We begin with a derivation of the Navier-Stokes equation which applies to constant density viscous fluids. This is then converted into an equation for vortex dynamics from which we derive an extended form of the famous Helmholtz equation. We then use that to derive the well known properties of vortex motion such as the so called Magnus force, and discuss the experiment Rayfield and Reif used to measure the circulation and core diameter of quantized vortices in superfluid helium.

## **Table of Contents**

## **FRONT COVER**

MKS Units	Front cover-2
Dimensions	Front cover-2
Powers of 10	Front cover-2

### TABLE OF CONTENTS

## CHAPTER 1 INTRODUCTION TO CALCULUS

Limiting Process Cal 1-2
The Uncertainty Principle
Calculus Definition of Velocity
Acceleration Cal 1-7 Components Cal 1-7
Integration
New Functions Cal 1-15
LogarithmsCal 1-15 The Exponential FunctionCal 1-16 Exponents to the Base 10Cal 1-16 The Exponential Function y <sup>x</sup> Cal 1-16 Euler's Number e = 2.7183Cal 1-17
Differentiation and Integration Cal 1-18 A Fast Way to go Back and Forth Cal 1-20 Constant Acceleration Formulas Cal 1-20 In Three Dimensions Cal 1-22
More on Differentiation
Integration Formulas Cal 1-27 Derivative of the Exponential Function Cal 1-28 Integral of the Exponential Function Cal 1-29
Derivative as the Slope of a Curve Cal 1-30 Negative Slope Cal 1-31
The Exponential Decay Cal 1-32 Muon Lifetime Cal 1-32 Half Life Cal 1-33 Measuring the Time Constant Cal 1-34
The Sine and Cosine Functions

## CHAPTER 2 SECOND DERIVATIVES AND THE ONE DIMENSIONAL WAVE EQUATION

The Second Derivative	Cal	2-2
Example 1	Cal	2-2
Geometrical Interpretation	Cal	2-3
Curvature	Cal	2-4
Curve Fitting and Boat Lofting	Cal	2-5
The Binomial Expansion	Cal	2-6
The Taylor Series Expansion	Cal	2-7
The Constant Acceleration Formulas	Cal	2-9
The Wave Equation	)al 2 )al 2 )al 2 )al 2 )al 2 )al 2	2-10 2-10 2-12 2-14 2-15 2-17

## CHAPTER 3 THE GRADIENT

Two views of the gradient	Cal 3-2
Calculating the Electric Field	Cal 3-4
Interpretation	Cal 3-6
The Gradient OPERATOR	Cal 3-7
The Parallel Plate Capacitor	Cal 3-8
Voltage Inside a Conductor	Cal 3-9
Electric Field of a Point Charge	Cal 3-10
Gradient in Cartesian Coordinates	Cal 3-12
Gradient in Cylindrical Coordinates	Cal 3-14
Gradient in Spherical Coordinates	Cal 3-16
Summary of Gradient Formulas	Cal 3-18
Cartesian Coordinates	Cal 3-18
Cylindrical Coordinates	Cal 3-18
Spherical Coordinates	Cal 3-18
Examples	Cal 3-18
Electric Field of a Point Charge	Cal 3-18
Electric Field of a Line Charge	Cal 3-19
The Coaxial Cable	Cal 3-21

## CH 3 VIEW 2 THE GRADIENT FROM A GEOMETRICAL PERSPECTIVE

Slope in Two Dimensions	Cal	3-23
The Gradient	Cal	3-25
Gradient as a Vector Field	Cal	3-28

### CH 3 VIEW 3

## CHAPTER 4 THE OPERATOR $\nabla^2 \equiv \vec{\nabla} \cdot \vec{\nabla}$

Fluid theory	Cal	4-1
Schrödinger's Equation	Cal	4-2
The Formulary	Cal	4-2
$ abla^2 $ in Cartesian Coordinates	Cal	4-3
$ abla^2$ in Spherical Polar Coordinates	Cal	4-3
Newtonian Fluids	Cal	4-4
Viscous Force on a Fluid Element	Cal	4-5
Viscous Force for 3D Flows Viscous Force in	Cal	4-6
Cylindrical Coordinates Measuring the Viscosity Coefficient	Cal Cal	4-7 4-9

## APPENDIX THE OPERATOR $\nabla^2$ in spherical polar coordinates

Spherical Polar Coordinates	Cal 4-12
Derivatives of r	Cal 4-13
Derivatives of $\hat{\theta}$	Cal 4-14
Derivatives of $\hat{\phi}$	Cal 4-14
Summary of Derivatives	Cal 4-16
Calculation of $\nabla^2 f$	Cal 4-16

## CHAPTER 5 INTRODUCTION TO COMPLEX VARIABLES

A Road Map	Cal	5-1
Introduction	Cal	5-1
Imaginary Numbers	Cal	5-2
Complex Numbers	Cal	5-2
Exponential Form of Complex Number Small Angle Approximations The Complex Conjugate Z <sup>*</sup>	Cal Cal Cal	5-3 5-4 5-6
Differential Equations for R, L, C Circuits . The RC Circuit An Aside on Labeling Voltages Solving the RC Circuit Equation The LC Circuit	Cal Cal Cal Cal Cal	5-6 5-6 5-7 5-8 5-8
A Faster Way to Find Real Solutions C	Cal 5	-10
The RLC Circuit C The Easy Way C	Cal 5 Cal 5	-11 -12
Impedance	Cal 5 Cal 5 Cal 5	-15 -18 -19
Transients       C         Particular Solution       C         Transient Solutions       C         Combined Solutions       C         Solutions of the 1D Wave Equation       C	Cal 5 Cal 5 Cal 5 Cal 5 Cal 5	-22 -22 -22 -23 -23

#### CHAPTER 6 INTRODUCTION TO SCHRÖDINGER'S EQUATION

Schrödinger's Wave Equation Cal 6-2
Potential Energy & Schrödinger's EquationCal 6-6
The Hydrogen Atom Cal 6-6
Interpretation of Solutions to Schrödinger's
Equation Cal 6-9
Normalization Cal 6-10
The Dirac EquationCal 6-12
Appendix I – Evaluation of a Normalization
Integral Cal 6-13
Appendix II – Schrödinger's Equation Applied to the Hydrogen Atom
The Hydrogen Atom Cal 6-14
The Second Energy Level Cal 6-16
Non Spherically Symmetric Solutions. Cal 6-18

## CHAPTER 7 DIVERGENCE

The Divergence	Cal 7-2
Electric Field of a Point Charge	Cal 7-6
The $\delta$ Function	Cal 7-8
Divergence Free Fields 0	Cal 7-10
Appendix - Flux Equation Derivation C	Cal 7-11

## CHAPTER 8 CURL

ABOUT THE CURL	Cal 8-1
Introduction to the curl	Cal 8-2
Stokes' Law	Cal 8-4
Ampere's Law	Cal 8-7
Curl of the Magnetic Field of a Wire C	al 8-10
Curl in Cylindrical Coordinates C	al 8-11;
Curl of the Magnetic Field of a Wire C Outside the wire C Inside the Wire	al 8-12 al 8-12 al 8-13

## CHAPTER 9 ELECTROMAGNETIC WAVES

Vector Identities	Cal	9-2
Derivation of the Wave Equation	Cal	9-4
Plane Wave Solution	Cal	9-6
Three Dimensional Wave Equation	Cal	9-7
Appendix – Order of partial differentiation	Cal	9-8

### CHAPTER 10 CONSERVATION OF ELECTRIC CHARGE

The Continuity Equation	Cal	10-2
From Maxwell's Equations	Cal	10-2
Integral Form of Continuity Equation	Cal	10-3

#### CHAPTER 11 SCALAR AND VECTOR POTENTIALS

The Vector Potential	. Cal 11-2
Wave Equations for $\phi$ and $\vec{A}$	Cal 11-4
Summary	. Cal 11-6

## CHAPTER 12 VORTICITY

DIVERGENCE FREE FIELDS	Cal	12-2
The Vorticity Field	Cal	12-2
Potential Flow	Cal	12-3
Examples of Potential Flow	Cal	12-4
Potential Flow in a Sealed Container	Cal	12-4
Potential Flow in a Straight Pipe	Cal	12-5
Superfluids	Cal	12-6
Vorticity as a Source of Fluid Motion	Cal	12-7
Picturing Vorticity	Cal	12-8
Solid Body Rotation	Cal	12-9
Vortex Core	Cal <sup>-</sup>	12-10
Stokes' Law Revisited Total Circulation and	Cal 1	12-11
Density of Circulation	Cal 1	12-11
Velocity Field of a Rotating Shaft	Cal 1	12-12
Wheel on Fixed Axle	Cal 1	12-12
A Conservation Law for Vorticity	Cal 1	12-13
Circulation of a Vortex	Cal 1	12-14
Quantum Vortices	Cal 1	12-15
Circulation of a Quantum Vortex	Cal 1	12-15
Rotating Bucket of Superfluid Helium	Cal 1	12-16
Bose-Einstein Condensates	Cal 1	12-18
The Vorticity Field	Cal <sup>-</sup>	12-18
Helmholtz Theorem	Cal 1	12-19
The Two Dimensional "Vortex Ring" .	Cal 1	12-19
The Circular Vortex Ring	Cal 1	12-20
Smoke Rings	Cal 1	12-20
Creating the Smoke Ring	Cal 1	12-21

## CHAPTER 13 INTRODUCTION TO FLUID DYNAMICS

The Navier-Stokes Equation Rate of Change of Momentum Einstein Summation Convention Mass Continuity Equation Rate of Change of Momentum	Ca Ca Ca Ca	ul 13-2 ul 13-2 ul 13-5 ul 13-5
when Mass is Conserved	Ca	ıl 13-6
Newton's Second Law	Ca	ıl 13-6
Bernoulli's Equation	Ca	al 13-8
Applies Along a Streamline	Ca	al 13-9
The Viscosity Term	. Cal	13-10
The Helmholtz Theorem	. Cal	13-11
Equation for Vorticity	. Cal	13-11
Non Potential Forces	. Cal	13-11
A Vector Identity for a Moving Circuit The Integral Form of the Vortex Dynamics Equation	. Cal . Cal	13-12 13-15
The Helmholtz Theorem Extended Helmholtz Theorem The Rayfield-Reif Experiment Motion of Charged Vortex Rings Conservation of Energy	. Cal . Cal . Cal . Cal . Cal . Cal	13-16 13-16 13-16 13-18 13-19
Quantized Circulation $\kappa = h/m_{He}$	. Cal	13-20
The Magnus Equation	. Cal	13-20
Impulse of a Vortex Ring	. Cal	13-23
The Airplane Wing	. Cal	13-24
The Magnus Lift Force	. Cal	13-25
The Magnus Force & Fluid Vortices	. Cal	13-26

## CH 13 APP 1 COMPONENT NOTATION

The Summation ConventionCa	ıl 13	A1	-2
The Dot Product and $\delta_{ij}$ Ca	ıl 13	A1	-2
The Cross Product and $\hat{\epsilon}_{iik}$ Ca	il 13	A1	-3
Handling Multiple Cross Products . Ca	ıl 13	A1	-5
Proof of the $\epsilon$ IdentityCa	ıl 13	A1	-6

## CH 13 APP 2 VORTEX CURRENTS

Conserved Two
Dimensional CurrentsCal 13 A2-2
Continuity Equation for Vorticity Cal 13 A2-3
A Single Vortex LineCal 13 A2-4
Center of Mass Motion Cal 13 A2-5
Magnus Formula for Curved Vortices Cal 13 A2-7
Creation of VorticityCal 13 A2-9
Energy Dissipation in Fluid Flow Cal 13 A2-10

## FORMULARY

Cylindrical Coordinates	Formulary-2
Divergence	Formulary-2
Gradient	Formulary-2
Curl	Formulary-2
Laplacian	Formulary-2
Laplacian of a vector	Formulary-2
Components of $(\vec{A} \cdot \vec{\nabla})\vec{B}$	Formulary-2
Spherical Polar Coordinates	Formulary-3
Divergence	Formulary-3
Gradient	Formulary-3
Curl	Formulary3
Laplacian	Formulary-3
Laplacian of a vector	Formulary-3
Components of $(\vec{A} \cdot \vec{\nabla})\vec{B}$	Formulary-3
Vector Identities	Formulary-4
Integral Formulas	Formulary-5
Working With Cross products	Formulary-6
The cross product	Formulary-6
Product of <b>E</b> 's	Formulary-6
Example of use	Formulary-6
Tensor Formulas	Formulary-7
Definition	Formulary-7
Formulas	Formulary-7
Div. of Tensor (Cylindrical Coord.)	Formulary-7
Div. of Tensor (Spherical Coord.)	Formulary-7
Short Table of Integrals	Formulary-8
Series Expansions	Formulary-9
The binomial expansion	Formulary-9
Taylor series expansion	Formulary-9
Sing and cosing	Formulary-0
Evonential	Formulary 0
	r onnulai y-9

## INDEX

## **BACK COVER**

Physical Constants in CGS Units......Back cover-1 Conversion Factors.....Back cover-1

## Index

## Symbols

- $ec{
  abla} imes$ , del cross, curl chapter on Cal 8-1
- $^{*}ec{
  abla}\cdot$  , divergence, chapter on Cal 7-1
- $abla^2$ , del squared, chapter on Cal 4-1

 $\delta_{ij}$  and  $\epsilon_{ijk}$  , appendix on Cal 13 A1-1

 $\vec{\nabla} f$  , gradient operation, chapter on Cal 3-1

## A

Acceleration Calculus definition of Cal 1-7 Component equations Cal 1-8 Vector equation Cal 1-7 Constant acceleration formulas Calculus derivation Cal 1-20 In three dimensions Cal 1-22 Adiabatic expansion Sound waves Cal 2-18 Adobe Illustrator™ Cal 2-6 Airplane wing Magnus equation Cal 13-24 Magnus lift force Cal 13-24 Wing vortex Cal 13-24 Allowed standing wave patterns, hydrogen Cal 6-7 Ampere's law In differential form Cal 8-3 Derivation of Cal 8-7 Amplitude Of a sine wave Cal 1-37 Angular momentum Quantum vortices as giant Bohr atom Cal 12-15 Schrödinger's equation solutions Cal 6-18 Area Related to integration Cal 1-11 Under the curve Cal 1-12 Atoms Standing wave patterns in hydrogen Cal 6-8

## B

Bernoulli's equation Applies along a streamline Cal 13-9 Derivation from Navier-Stokes equation Cal 13-8 For potential flow Cal 13-9 Gradient of hydrodynamic voltage Cal 13-9 Hydrodynamic voltage Cal 13-9 Besier curves Cal 2-6 Binomial expansion Cal 1-23 Derivation of Cal 2-6, Form.-9 Boat lofting Cal 2-5 Bohr atom Quantum vortices Cal 12-15 Bohr radius Cal 6-7 Born, Max Interpretation of solutions to Schrödinger's Eq. Cal 6-9

## С

Calculation of integrals Cal 1-11 Calculus And the uncertainty principle Cal 1-2 Calculating integrals Cal 1-11 Calculus in physics Cal 1-1 Chain rule Cal 1-25 Definition of acceleration Cal 1-7 Component equations Cal 1-8 Vector equation Cal 1-7 Definition of velocity Cal 1-5 Component equations Cal 1-8 Vector equation Cal 1-6 Derivation of constant acc. formulas Cal 1-20 In three dimensions Cal 1-22 Limiting process Cal 1-2, Cal 1-5 Vector equation for Cal 1-5 Special chapter on Cal 1-1 Calculus in Physics i, Cal 2-1, Cal 3-1, Cal 4-1, Cal 5-1, Cal 6-1, Cal 7-1, Cal 8-1, Cal 9-1, Cal 10-1, Cal 11-1, Form.-1 Capacitance Dimensions of Front cover-2 Cartesian coordinates Del squared Cal 4-3 Gradient Cal 3-12 Right hand rule Cal 3-12 Unit vectors Cal 3-12 Center of mass Vortex line motion Cal 13 A2-4 Center of mass motion Conserved vortex current Cal 13 A2-5 Vortex line Cal 13 A2-5 CGS units Back cover-1 Chain rule Cal 1-25 Proving it (almost) Cal 1-26 Remembering it Cal 1-25 Charge Electric Conservation due to Maxwell's equations Cal 10-1 Continuity equation for Cal 10-2 Choice of gauge Cal 11-4

Circuits Driven LRC circuit Cal 5-19 Resonance in Cal 5-21 Impedance Cal 5-15 Impedance formulas Cal 5-18 LC circuit Ringing like a bell Cal 5-11 RC circuit Solving with complex numbers Cal 5-8 RLC circuit Decaying oscillation Cal 5-10 Differential equation for Cal 5-11 Solution using complex variables Cal 5-12 Transient solutions Cal 5-22 Circulation Density of Cal 12-11 For roller bearings Cal 12-13 For vortex sheet Cal 12-13 For wheel on fixed axle Cal 12-13 Flux of vorticity in flow tube Cal 12-18 Measurement of quantized value Cal 13-20 Of a quantum vortex Cal 12-15 Of a Vortex Cal 12-14 Unchanged by local non potential forces Cal 13-17 Total Cal 12-11 Circulation, total Of rotating shaft Cal 12-12 Classical theory of electromagnetism Cal 11-3 Coaxial cable Example of gradient in Cyl. Coord. Cal 3-21 Voltage in Cal 3-21 Coefficient of viscosity Cal 4-4 Measuring Cal 4-9 Experimental formula for Cal 4-11 Second viscosity coefficient Cal 4-6 Complex analysis Of driven LRC circuit Cal 5-19 Complex conjugate Cal 5-6 Of Schrödinger wave function Cal 6-9 Complex numbers Cal 5-2 Analogy to coordinate vector Cal 5-2 As a complex exponential Cal 5-5 Exponential form Cal 5-3 Plotting Cal 5-2 Real part, imaginary part Cal 5-2 Solving the LC circuit Cal 5-9 Solving the RLC circuit Cal 5-12 Transient solutions Cal 5-22

Complex variables Chapter on Cal 5-1 Complex numbers Cal 5-2 Exponential function Series expansion Cal 5-4 Fast way to find real solutions Cal 5-10 Imaginary numbers Cal 5-2 Impedance Cal 5-15 Driven LRC circuit Cal 5-19 Formulas for Cal 5-18 Why Schrödinger's equation is complex Cal 6-5 Component notation Appendix on Cal 13 A1-1 Components, vector Cal 1-7 In rotated coordinate system Cal 3-26 Compressible fluids Continuity equation Integral form Cal 10-3 Compressional waves on spring Cal 2-15 Computers Why they are so good at integration Cal 1-12 Conjugate, complex Cal 5-6 Conservation law For vorticity Cal 12-13 Related field Cal 10-3 Conservation of Electric charge Chapter on Cal 10-1 Conservation of charge Related to electric fields Cal 10-3 Conservation of energy In vortex ring motion Cal 13-19 Related to gravity Cal 10-3 Conservative field Fluid flow Cal 12-3 Conservative force And Faraday's law Cal 11-2 Conserved current Vortex, intuitive discussion of Cal 13 A2-2 Constant acceleration formulas Calculus derivation Cal 1-20 In three dimensions Cal 1-22 Second derivative Cal 2-9 Constant, integral of Cal 1-13 Continuity equation Derivation from Maxwell's equations Cal 10-2 For compressible fluids Differential form Cal 10-2 Integral form Cal 10-3 For electric charge and current Cal 10-2 For flow of mass Cal 13-5 Formula for Cal 13-6 For flow of vorticity Cal 13 A2-1 Derivation of Cal 13 A2-3 Conversion factors Back cover-1

Coordinate system Cartesian Cal 3-12 Del squared Cal 4-3 Unit vectors Cal 3-12 Cvlindrical Cal 3-14 Unit vectors Cal 3-14 Spherical Cal 3-16 Derivation of del squared in Cal 4-12 Unit vectors Cal 3-16 Coordinate system rotated Components of a vector in Cal 3-26 Core Vortex core structure Cal 12-10 Analogy to magnetic field Cal 12-10 Cosine function Amplitude of Cal 1-37 As function of complex exponential Cal 5-5 Definition of Cal 1-35 Derivative of Cal 1-38 Series expansion Cal 5-4 Coulomb potential in Schrödinger's equation Cal 6-6 Creating a smoke ring Cal 12-21 Creation of vorticity Cal 13 A2-9 Cross product Multiple, easy way to handle Cal 13 A1-5 Relation to curl Cal 8-2 Use of epsilon i.j.k Cal 13 A1-3 Working with Form.-6 Curl Chapter on Cal 8-1 Definition of Cal 8-2 In cylindrical coordinates Cal 8-11 Vorticity calculation Cal 12-9 Introduction to Cal 8-2 Line integral shrunk down Cal 8-1, Cal 8-3 Of a divergence = 0 Cal 11-1 Of magnetic field of wire Cal 8-10 Calculating the curl Cal 8-12 Of solid body rotation velocity field Cal 12-10 Of vector potential Cal 11-4 Gauge invarience Cal 11-4 Of vortex velocity field Cal 12-9 Of vorticity Viscosity term in Navier-Stokes equation Cal 13-10 Curl and divergence Uniquely determined field Cal 12-2 Curl theorem Called Stokes law Cal 8-3 Current, vortex Appendix on Cal 13 A2-1 Conserved, intuitive discussion of Cal 13 A2-2 Continuity equation for Cal 13 A2-3

Curvature, radius of Definition Cal 2-4 Rope waves Cal 2-10 Second derivative Cal 2-4 Curve Area under, integral of Cal 1-12 Besier (Adobe Illustrator) Cal 2-6 Slope as derivative Cal 1-30 That increases linearly, integral of Cal 1-13 Velocity, area under Cal 1-12 Curve fitting Cal 2-5 Cylindrical coordinates Curl in Cal 8-11 Curl in solid body rotation Cal 12-9 Curl of magnetic field of wire Cal 8-12 Div, grad, curl, del squared, A dot del B Form.-2 Gradient in Cal 3-14 Radial component Cal 3-14 Theta component Cal 3-15 Unit vectors Cal 3-14 Viscous force in Cal 4-7

#### **D** De Broglie

Schrödinger's equation Cal 6-2 Debye, on electron waves Cal 6-1 Decay Exponential decay Cal 1-32 Decaying oscillation RLC circuit Cal 5-10 Definite integral Compared to indefinite integrals Cal 1-14 Defining new functions Cal 1-15 Introduction to Cal 1-11 Of velocity Cal 1-11 Process of integrating Cal 1-13 Del Relation to curl Cal 8-2 Del - gradient operator Cal 3-7 Del cross; curl Chapter on Cal 8-1 Del squared Chapter on Cal 4-1 In Cartesian coordinates Cal 4-3 In spherical polar coordinates Derivation of Cal 4-12 Spherical harmonics Cal 6-18 Relation to curl Cal 8-2 Relation to potential flow Cal 12-4 Schrödinger's equation Applied to hydrogen atom Cal 6-14 Schrödinger's equation Cal 4-2 Viscous force Cal 4-1 For 3D flows Cal 4-6 In cylindrical coordinates Cal 4-7

Delta function Definition of Cal 7-8 In three dimensions Cal 7-8 Used in Gauss' law Cal 7-9 Delta i.i Handling multiple cross products Cal 13 A1-5 Used in dot product Cal 13 A1-2 Delta i,j and epsilon i,j,k Appendix on Cal 13 A1-1 Density of circulation Stokes' law Cal 12-11 Derivative As a limiting process Cal 1-6, Cal 1-18, Cal 1-23, Cal 1-28, Cal 1-30 As the Slope of a Curve Cal 1-30 Constants come outside Cal 1-24 Negative slope Cal 1-31 Of exponential function e to the x Cal 1-28 Of exponential function e to the ax Cal 1-29 Of function x to the n'th power Cal 1-24 Of sine function Cal 1-38 Partial Cal 5-24 Second Chapter on Cal 2-1 Constant acceleration formulas Cal 2-9 Third, boat lofting Cal 2-5 Derivative, partial Order of, appendix on Cal 9-8 Derivative, second Cal 2-2 Geometrical interpretation Cal 2-3 Of a sine wave Cal 2-2 Differential equation Fast way to find real solutions Cal 5-10 For LC circuit Solving with complex numbers Cal 5-8 For LRC circuit Transient solutions Cal 5-22 For R, L, and C circuits Cal 5-6 Homogenous Cal 5-9 To integral equation Cal 3-4 Differentiation. See also Derivative Chain rule Cal 1-25 More on Cal 1-23 Differentiation and integration As inverse operations Cal 1-18 Velocity and position Cal 1-18 Fast way to go back and forth Cal 1-20 Position as integral of velocity Cal 1-20 Velocity as derivative of position Cal 1-20

Dimensions of Capacitance Front cover-2 Electric charge Front cover-2 Electric potential Front cover-2 Electric resistance Front cover-2 Energy Front cover-2 Force Front cover-2 Frequency Front cover-2 Inductance Front cover-2 Magnetic field Front cover-2 Magnetic flux Front cover-2 Power Front cover-2 Pressure Front cover-2 Dirac equation Discussion of Cal 6-12 Divergence Chapter on Cal 7-1 Relation to curl Cal 8-2 Shrinking the surface integral Cal 7-2 Theorem Cal 7-5 Handling a point charge Cal 7-7 Relation to curl Cal 8-3 Divergence and curl Surface & line integrals shrunken Cal 7-1, Cal 7-2 Uniquely determined field Cal 12-2 Divergence and gradient compared Cal 7-5 Divergence free fields Cal 7-10 Dot product Relation to curl Cal 8-2 Use of delta i,j Cal 13 A1-2 Driven LRC circuit Cal 5-19

## E

Einstein Summation convention Cal 13-5 Electric and magnetic fields In terms of scalar and vector potentials Cal 11-3 Electric charge Conservation of Consequence of Maxwell's equations Cal 10-1 Continuity equation for Cal 10-2 Dimensions of Front cover-2 Electric field Gradient of voltage Cal 3-3 Equation for Cal 3-7 Field of point charge Cal 3-10 Interpretation Cal 3-6 In terms of scalar & vector potentials Cal 11-3 Of a line charge Using gradient in cylind. coord. Cal 3-19 Of a point charge Using gradient in spherical coord. Cal 3-18 Wave equation for With sources Cal 11-6 Electric potential Dimensions of Front cover-2 Plotting experiment Cal 3-2 Related to fluid flows Cal 12-3 Electric resistance Dimensions of Front cover-2 Electromagnetic waves Chapter on wave equation Cal 9-1 Electromagnetism Classical theory of Cal 11-3 Electron In Standard model of elementary particles Cal 7-7 Point particle? Cal 7-7 Energy Dimensions of Front cover-2 Energy levels, hydrogen Calculation of lowest Cal 6-15 Lowest two from Schrödinger's equation Cal 6-7 Epsilon i, j, k Use in cross product Cal 13 A1-3 Handling multiple cross products Cal 13 A1-5 Epsilon i,j,k and delta i,j Appendix on Cal 13 A1-1

Equation Continuity For electric charge and current Cal 10-2 Extended Helmholtz equation Cal 13-15 Magnus Cal 13-21 Airplane wing Cal 13-24 Maxwell's Derivation of the wave equation Cal 9-4 Vector identities for Cal 9-2 Navier-Stokes Cal 13-2 Nonlinear effects Cal 13-7 Schrödinger's. See Schrödinger wave equation Vector Components with derivatives Cal 1-7 Vortex dynamics equation Cal 13-12 Wave, one dimensional Cal 2-1 General form of Cal 2-14 Solutions using complex variables Cal 5-24 Wave, relativistic Dirac's Cal 6-12 For zero rest mass particles Cal 6-2 Particles with rest mass Cal 6-3 Schrödinger's Cal 6-3 Euler's number e = 2.7183... Cal 1-17 Expansion, binomial Cal 1-23 Derivation of Cal 2-6, Form.-9 Expansion, series Exponential function in complex variables Cal 5-4 Sin and cosine Cal 5-4 Taylor series Cal 2-7 Experiments II Potential plotting Cal 3-2 Exponential decay Cal 1-32 Exponential form complex number Cal 5-3 Exponential function As function of sin and cos Cal 5-5 Derivative of Cal 1-28 Exponential decay Cal 1-32 Indefinite integral of Cal 1-29 Integral of Cal 1-29 Introduction to Cal 1-16 Inverse of the logarithm Cal 1-16 Series expansion Cal 1-28 y to the x power Cal 1-16 Extended Helmholtz's theorem Cal 13-15 Discussion of Cal 13-16

## F

Fall line. See Gradient: Of voltage: Interpretation As a field line Cal 3-23 Faraday's law In terms of the vector potential Cal 11-3 Non potential field Cal 12-3 Feynman Cal 7-7 Quantized vortices Cal 12-16 Parabolic surface of rotating helium Cal 12-6 Field Divergence free Cal 7-10 Plotting experiment Cal 3-2 Pressure field Cal 3-1 Scalar field Cal 3-7 Uniquely determined, conditions for Cal 12-2 Vector field Created by gradient Cal 3-1 Vorticity field Cal 12-18 Field lines And contour lines Cal 3-23 Two dimensional slope Cal 3-24 Fluid dynamics Introductory chapter on Cal 13-1 Vorticity Cal 12-1 Fluids Compressible Continuity equation for Cal 10-3 Laminar flow Cal 4-8 Newtonian, definition of Cal 4-4 Potential flow Cal 12-3 In a straight pipe Cal 12-5 Zero vorticity Cal 12-3 Solid body rotation Cal 12-9 Viscous force on Cal 4-5 Vorticity as a source of fluid motion Cal 12-7 Flux Of vorticity in flow tube Cal 12-18 Rate of change of through moving circuit Cal 13-15 Flux equation, derivation of Cal 7-11 Force Conservative forces And Faraday's law Cal 11-2 Dimensions of Front cover-2 Non potential In Navier-Stokes equation Cal 13-11 Viscous In cylindrical coordinates Cal 4-7 In pipe flow Cal 4-7 On a fluid element Cal 4-5 Formulary Form.-1 Discussion of Cal 4-2 Fractal geometry Cal 3-23 Frequency Dimensions of Front cover-2

Functions delta i,j and epsilon i,j,k *Appendix on Cal 13 A1-1* Functions obtained from integration Cal 1-15 *Logarithms Cal 1-15* 

## G

Gamma Speed of sound Cal 2-18 Gauge invariance Choice of vector potential divergence Cal 11-4 Gauge invariant theory Cal 11-4 Gauss' law Derived from differential equation Cal 7-7 Electric field of point charge Using delta function Cal 7-9 Geometrical interpretation Of Gradient Cal 3-4, Cal 3-22 Equations for Cal 3-25 Of second derivative Cal 2-3 Geometry, fractal Cal 3-23 Gibbs, Willard; gradient notation Cal 3-7 Gradient A summary of gradient formulas Cal 3-18 As a vector field Cal 3-28 Chapter on Cal 3-1 From a Geometrical Perspective Cal 3-4, Cal 3-22 Equations for Cal 3-25 In Cartesian coordinates Cal 3-12 In cylindrical coordinates Cal 3-14 Coaxial cable Cal 3-21 Electric field of line charge Cal 3-19 Radial component Cal 3-14 Theta component Cal 3-15 In spherical coordinates Phi component Cal 3-17 Theta component Cal 3-17 Of pressure Cal 3-29 Of voltage Cal 3-3 Field of point charge Cal 3-10 Interpretation Cal 3-6 Parallel plate capacitor Cal 3-8 Voltage inside conductor Cal 3-9 Operator "del" Cal 3-7 Relation to curl Cal 8-2 Vector and scalar fields Cal 3-1 Gradient vector In three dimensions Cal 3-28 Steepest slope Cal 3-25 Transformation of Cal 3-25 Gravity Quantum theory of Cal 7-7 Gyroscope like behavior Of vortex line due to non potential force Cal 13-18 Gyroscopes Superfluid Cal 12-17

## Η

Half-life In exponential decay Cal 1-33 Of muons, exponential decay Cal 1-33 Heisenberg, Werner Cal 1-2 Helium, superfluid Cal 12-6 Helmholtz theorem Application to smoke rings Cal 12-21 Derivation from Navier-Stokes equation Cal 13-11 From extended Helmholtz theorem Cal 13-16 Introduction to Cal 12-19 Helmholtz theorem extended Discussion of Cal 13-16 Including non potential forces Cal 13-15 Homogeneous solution for RLC equation Cal 5-23 Homogenous differential equation Cal 5-9, Cal 5-23 Hydrodynamic voltage Gradient of and Bernoulli's equation Cal 13-9 Hydrogen atom Bohr radius Cal 6-7 Schrödinger's equation for Cal 6-6 Schrödinger's equation solutions Cal 6-7 Lowest two energy levels Cal 6-7 Non spherically symmetric Cal 6-18 Spherical harmonics Cal 6-18 Standing wave patterns in Cal 6-8 Hydrogen wave patterns Lowest energy ones Cal 6-8

## I

Illustrator™, Adobe Cal 2-6 Imaginary numbers Cal 5-2 Impedance Cal 5-15 Formulas for Cal 5-18 Impulse Of a vortex ring Cal 13-23 Impulse equation Cal 13-23 Indefinite integral Definition of Cal 1-14 Of exponential function Cal 1-29 Inductance Dimensions of Front cover-2 Infinities in the gravitational interaction String theory Cal 7-7 Instantaneous velocity And the uncertainty principle Cal 1-2 Calculus definition of Cal 1-5

Integral As a sum Cal 1-10 Calculating them Cal 1-11 Definite, introduction to Cal 1-11 Formula for integrating x to n'th power Cal 1-14, Cal 1-27 Indefinite, definition of Cal 1-14 Of 1/x, the logarithm Cal 1-15 Of a constant Cal 1-13 Of a curve that increases linearly Cal 1-13 Of a velocity curve Cal 1-12 Of exponential function e to the ax Cal 1-29 Of the velocity vector Cal 1-10 As area under curve Cal 1-12 Of x to n'th power Indefinite integral Cal 1-27 Integral formulas Many of them Form.-5 Integral, line Becomes curl for infinitesimal paths Cal 8-3 Integral sign Cal 1-10 Integral, surface Shrinking for divergence Cal 7-2 Integral to differential equations Cal 3-4 Integration Equivalent to finding area Cal 1-11 Introduction to Cal 1-8 Introduction to finding areas under curves Cal 1-13 Why computers do it so well Cal 1-12 Integration and differentiation As inverse operations Cal 1-18 Fast way to go back and forth Cal 1-20 Position as integral of velocity Cal 1-20 Velocity as derivative of position Cal 1-20 Integration formulas Cal 1-27 Intensity Of wave function Cal 6-9 Interpretation of solutions to Schrödinger's Eq. Cal 6-9 Interval, evaluating variables over Cal 1-10

## L

Laminar flow Cal 4-8, Cal 7-10 Landau, Lev Superfluid helium Cal 12-6 Landau's prediction for Cal 12-6 Laplacian Relation to curl Cal 8-2 Laplacian (del squared) Chapter on Cal 4-1 Relation to potential flow Cal 12-4 LC circuit Ringing like a bell Cal 5-11 Leibnitz Cal 1-2 Leptons Standard model of elementary particles Cal 7-7 Lifetime Muon, exponential decay Cal 1-32 Light Speed of light From one dimensional wave equation Cal 9-7 Structure of electromagnetic wave Cal 9-1, Cal 9-6 Limiting process Cal 1-2 Definition of derivative Cal 1-30 In calculus Cal 1-5 Introduction to derivative Cal 1-6 With strobe photographs Cal 1-3 Line charge, electric field of Calculated using calculus In cylindrical coordinates Cal 3-19 Line integral Becomes curl for infinitesimal paths Cal 8-3 Localized non potential force Effect on vortex motion Cal 13-17 Lofting, boat Cal 2-5 Logarithms Integral of 1/x Cal 1-15 Introduction to Cal 1-15 Inverse of exponential function Cal 1-16 LRC circuit. See RLC circuit LRC circuit, ringing like a bell Cal 5-11

## М

Magnetic and electric fields In terms of scalar and vector potentials Cal 11-3 Magnetic field Analogous to vorticity in fluids Cal 12-7 Of a straight wire Calculating curl of Cal 8-12 Curl of Cal 8-10 Wave equation for With sources Cal 11-6 Magnetic flux Dimensions of Front cover-2 Magnus equation Airplane wing Cal 13-24 Relative motion of vortex line and fluid particles Cal 13-20 The equation Cal 13-21 Magnus formula Exact for curved vortices Cal 13 A2-1 Magnus lift force Cal 13-25 On fluid core vortices - a pseudo force Cal 13-26 Mass Continuity equation for flow of Cal 13-5 Maxwell's equations All forms of Cal 11-6 Conservation of electric charge Cal 10-1 Derivation of the wave equation Cal 9-4 In differential form Cal 8-9 In terms of scalar and vector potentials Cal 11-3 Introducing vector potential into Cal 11-3 One dimensional wave equation Gives speed of light Cal 9-7 Plane wave solution Cal 9-6 Relativistic wave equation for photons Cal 6-3 Vector identities for Cal 9-2 Vector potential in Cal 11-2 Measurement limitation Due to uncertainty principle Cal 1-2 Measurement of quantized circulation Cal 13-20 Measuring time constant from graph Cal 1-34 MKS units Front cover-2 Modulus Spring Cal 2-15 Momentum of fluid particles Navier-Stokes equation Cal 13-2 Motion Of charged vortex rings Cal 13-18 Of vortex line, relative directions Cal 13-22 Moving circuit Vector Identity for Cal 13-12 Multiple cross products Easy way to handle Cal 13 A1-5 Muon In Standard model of elementary particles Cal 7-7 Lifetime, exponential decay Cal 1-32

## N

Navier-Stokes equation Cal 13-2 As starting point for fluid theory Cal 13-7 Bernoulli equation derivation Cal 13-8 Derivation of Helmholtz's theorem from Cal 13-11 Final equation! Cal 13-7 Momentum of fluid particles Cal 13-2 Newton's second law for fluids Cal 13-2 Non potential forces in Cal 13-11 Nonlinear equation Cal 13-7 Rate of change of momentum Cal 13-2 Role of viscosity Cal 13-7 Viscosity term in Cal 13-10 Curl of vorticity Cal 13-10 Negative slope Cal 1-31 Neutrinos In Standard model of elementary particles Cal 7-7 New functions, obtained from integration Cal 1-15 Newtonian Fluids Definition of Cal 4-4 Newton's laws Second law For fluids, the Navier-Stokes equation Cal 13-2 Non potential field Cal 12-3 Non potential forces In extended Helmholtz theorem Cal 13-15 In Navier-Stokes equation Cal 13-11 Localized Causing sideways motion Cal 13-17 Ravfield-Reif experiment Cal 13-16 Nonlinear equation Navier-Stokes equation Cal 13-7 Normalization of wave function Cal 6-10

## 0

One dimensional wave equation Cal 2-1, Cal 2-14 Maxwell's equations Gives speed of light Cal 9-7 Solutions using complex variables Cal 5-24, Cal 5-25 Order of partial derivative Cal 9-8 Oscillation Decaying Cal 5-10

## P

Parabolic profile, pipe flow Cal 4-8 Parabolic surface, rotating fluid Superfluid helium Cal 12-6 Telescope mirror Cal 12-6 Parallel plate capacitor Example of voltage gradient Cal 3-8 Partial derivative Cal 5-24 Order of Appendix on Cal 9-8 Partial derivative operator Cal 8-2 Particular solution, driven RLC circuit Cal 5-22 Perpendicular components of flow Cal 12-2 Phi component Gradient in spherical coordinates Cal 3-17 Photons Relativistic wave equation for Cal 6-3 Physical constants In CGS units Back cover-1 In MKS units Front cover-2 Pipe flow Calculating viscous forces Cal 4-7 Measuring viscosity coefficient Cal 4-9 Parabolic profile Cal 4-8 Potential flow in Cal 12-5 Pressure force Cal 4-9 Viscous force formula Cal 4-8 Plane, tangent Cal 3-23 Plane wave Discussion of Cal 9-6 Solution for Maxwell's equations Cal 9-6 Plotting Experiment, electric potential Cal 3-2 Plywood model. See Gradient: Of voltage: Interpretation Point charge Divergence theorem Cal 7-7 Quantum electrodynamics Cal 7-7 Point particles Delta function Cal 7-8 Problems with gravity theory Cal 7-7 Standard model Cal 7-7 Postscript<sup>™</sup> language Cal 2-6

Potential, magnetic Wave equation for Cal 11-4 Potential, electric Wave equation for Cal 11-4 Potential energy Electric potential energy Electric field as gradient of Cal 3-4 Schrödinger's Equation Cal 6-6 Potential flow And the Laplacian (del squared) Cal 12-4 Bernoulli's equation in Cal 13-9 Definition of Cal 12-3 Examples of In a sealed container Cal 12-4 In a straight pipe Cal 12-5 Superfluids Cal 12-6 Zero curl, no vorticity Cal 12-3 Power Dimensions of Front cover-2 Power series. See Series expansions Powers of 10, names of Front cover-2 Prediction of motion Using calculus Cal 1-9 Pressure Dimensions of Front cover-2 Pressure field Cal 3-1 Pressure force As gradient of pressure Cal 3-29 In pipe flow Cal 4-9 Per unit volume Cal 3-30 Probability wave, Schrödinger's Equation Cal 6-9 Projectile motion And the uncertainty principle Cal 1-4 Calculus definition of velocity Cal 1-5 Pulse Formation of wave pulse Cal 2-14

## Q

Quantized angular momentum In hydrogen wave patterns Cal 6-8 Quantized circulation Measurement of Cal 13-20 Quantized vortex ring Rayfield-Reif experiment Cal 13-16 Quantum electrodynamics Feynman, Schwinger, and Tomonaga Cal 7-7 Point charges Cal 7-7 Quantum mechanics Concept of velocity Cal 1-4 Quantum theory Vector potential needed in Cal 11-3 Quantum theory of gravity Cal 7-7 Quantum vortices Cal 12-15 Core of Cal 12-16 Giant Bohr atom Cal 12-15 Number in rotating bucket Cal 12-16 Quarks Cal 7-7 In Standard model of elementary particles Cal 7-7

## R

Radial component Gradient in cylindrical coordinates Cal 3-14 Radian measure Cal 1-35 Radians to degrees Cal 5-4 Radius of curvature Definition Cal 2-4 Second derivative Cal 2-4 Rate of change of momentum Of fluid particles, Navier-Stokes equation When mass is conserved Cal 13-6 Rayfield-Reif experiment Cal 13-16 Creation of vorticity Cal 13 A2-9 Motion of charged vortex rings Cal 13-18 RC circuit Differential equation for Cal 5-6 Solving with complex numbers Cal 5-8 Labeling voltages Cal 5-7 Real part of complex number Cal 5-2 Relativistic physics Electromagnetic radiation, structure of Cal 9-1. Cal 9-6 Relativistic wave equation For zero rest mass particles Cal 6-2 Particles with rest mass Cal 6-3 Schrödinger's Cal 6-3 Resonance In driven RLC circuits Cal 5-21 Rest mass Non zero Relativistic wave equation for Cal 6-3 Zero Relativistic wave equation for Cal 6-2

**Right-hand rule** For Cartesian coordinates Cal 3-12 **RLC** circuit Decaying oscillation Cal 5-10 Differential equation for Cal 5-11 Driven LRC circuit Cal 5-19 Impedance Cal 5-15 Formulas for Cal 5-18 Labeling voltages Cal 5-7 Solution using complex variables Cal 5-12 Transient solutions Cal 5-22 Roller bearings For wheel on fixed axle, Stokes' law Cal 12-13 Rope Wave equation for Cal 2-10 Rotated coordinate system Components of a vector in Cal 3-26 Rotating bucket of superfluid helium Quantized vortices in Cal 12-16 Rotating shaft Total circulation of Cal 12-12 Velocity field of Cal 12-12

## S

Scalar and vector potentials *Chapter on Cal 11-1* Scalar field Cal 3-7 *Gradient gives vector field Cal 3-29 Pressure Cal 3-29* Scalar potential *And the electric field Cal 11-3 Relation to vector potential Cal 11-2 Wave equation for Cal 11-4 Coulomb gauge Cal 11-6 Gauge invariant form Cal 11-4 Wave gauge Cal 11-5* Schrödinger, Erwin Cal 6-1 Schrödinger's equation Allowed standing wave patterns, hydrogen Cal 6-Angular momentum in solutions Cal 6-18 Applied to the hydrogen atom Cal 6-14 Bohr radius Cal 6-7 Calculation of lowest energy level Cal 6-15 Chapter on Cal 6-1 Complex conjugate of wave function Cal 6-9 Coulomb potential Cal 6-6 Del squared in Cal 4-2 Felix Block story on Cal 6-1 For hydrogen atom Cal 6-6 Full three dimensional form Cal 6-6 Hydrogen atom solution Cal 6-6, Cal 6-7 Ideas that led to it Cal 6-2 Intensity of wave function Cal 6-9, Cal 6-10 Interpretation of solutions Cal 6-9 Lowest two energy levels Cal 6-7 Non spherically symmetric solutions Cal 6-18 Spherical harmonics Cal 6-18 Normalization of wave function Cal 6-10 Potential energy in Cal 6-6 Probability interpretation Cal 6-9 Second energy level Cal 6-16 Solutions of definite energy Cal 6-14 Solved for hydrogen atom Cal 6-14 Why it is complex Cal 6-5 Schrödinger's relativistic wave equation Cal 6-3 Two solutions Cal 6-4 Schwinger Cal 7-7 Second derivative Cal 2-2 Constant acceleration formulas Cal 2-9 Geometrical interpretation Cal 2-3 Of a sine wave Cal 2-2 Radius of curvature Cal 2-4 Second energy level, hydrogen Cal 6-16 Second viscosity coefficient Cal 4-6 Series expansions Cal 1-23 Binomial Cal 1-23 Exponential function Complex variables Cal 5-4 Exponential function e to the x Cal 1-28 Sine and cosine Cal 5-4 Taylor Cal 2-7 Sideways motion of vortex line Caused by localized non potential force Cal 13-17 Sine function Amplitude of Cal 1-37 Definition of Cal 1-35. Cal 1-36 Derivative of, derivation Cal 1-38 Series expansion Cal 5-4 Sine waves As function of complex exponential Cal 5-5 Second derivative Cal 2-2 Traveling wave Cal 2-14

Single vortex line Cal 13 A2-4 Singly connected surface, Stokes' law Cal 12-13 Slinky™ Compressional wave on Cal 2-15 Slope of a curve And contour maps Cal 3-23 As derivative Cal 1-30, Cal 3-23 Formula for Cal 1-30 In two dimensions Cal 3-23 Negative slope Cal 1-31 Steepest slope, gradient vector Cal 3-25 Smoke rings Cal 12-20 Approaching each other Cal 12-21 Creating Cal 12-21 Role of viscosity in Cal 12-21 Prediction of Helmholtz's theorem Cal 12-21 Stability of Cal 12-21 Titanium tetrachloride for Cal 12-20 Soap film analogy Stokes' law Cal 8-7 Solar neutrinos. See Neutrinos Solid body rotation Cal 12-9 Curl of velocity field Cal 12-10 Sound Speed, formula for Cal 2-20 Speed of air molecules Cal 2-21 Speed of, calculating Cal 2-17 Wave equation for Cal 2-17, Cal 2-20 Adiabatic expansion Cal 2-18 Sound waves, speed of Formula for Cal 2-21 Source of fields, conserved Cal 10-3 Source terms for wave equations Cal 11-6 Speed of Air molecules Cal 2-21 Sound, formula for Cal 2-20 Vortex rinas Circular rings Cal 12-20 Two dimensional rings Cal 12-19 Wave pulses On rope, calculus derivation Cal 2-13 Waves One dimensional wave equation Cal 5-24 Spherical coordinates Derivation of del squared in Cal 4-12 Div, grad, curl, del squared, A dot del B Form.-3 Gradient in Cal 3-16 Phi component Cal 3-17 Theta component Cal 3-17 Schrödinger's equation Cal 6-6 Unit vectors Cal 3-16 Derivative of changing unit vectors Cal 4-12 Spherical harmonics Cal 6-18

Spline fitting Cal 2-5 Spring Wave equation for Cal 2-17 Speed of wave Cal 2-17 Spring modulus Cal 2-15 Stability of smoke rings Cal 12-21 Standard model of elementary particles Cal 7-7 Leptons Cal 7-7 Electrons Cal 7-7 Muons Cal 7-7 Neutrinos Cal 7-7 Tau particle Cal 7-7 Quarks Cal 7-7 Standing waves Patterns in hydrogen Cal 6-8 From Schrödinger's equation Cal 6-7 Wave equation Cal 2-14 Stokes' law Applied to wheel on fixed axle Cal 12-13 Convert line to surface integral Cal 8-4 Derivation of Cal 8-4 Final result Cal 8-6 Introduction to Cal 8-3 Revisited Cal 12-11 Roller bearings Cal 12-13 Soap film analogy Cal 8-7 Total circulation and density of circulation Cal 12-11 Strain, definition of Wave equation Cal 2-15 Streamlines Bernoulli's equation Applies along a streamline, derivation of Cal 13-9 Stress Cal 4-4 Viscous Cal 4-4 String theory Cal 7-7 Vortex current tensor Cal 13 A2-1 Strobe photographs And the uncertainty principle Cal 1-2 Substantive derivative Cal 13-2 Summation Becoming an integral Cal 1-10 Of velocity vectors Cal 1-10 Summation convention, Einstein's Cal 13-5, Cal 13 A1-2 Superfluid gyroscope Cal 12-17 Superfluids Potential flow in Cal 12-6 Superfluid helium Cal 12-6 Feynman's prediction for Cal 12-6, Cal 12-16 Landau's prediction for Cal 12-6 Surface integral Shrinking for divergence Cal 7-2

## T

Tangent line Cal 3-23 Tangent plane Cal 3-23 Tau particle In Standard model of elementary particles Cal 7-7 Taylor series expansion Cal 2-7 Constant acceleration formulas Cal 2-9 Tensor formulas Form.-7 Theta component Gradient In cylindrical coordinates Cal 3-15 In spherical coordinates Cal 3-17 Third derivative, boat lofting Cal 2-5 Time constant Measuring from a graph Cal 1-34 Titanium tetrachloride for smoke rings Cal 12-20 Tomonaga Cal 7-7 Total circulation Of a vortex Cal 12-14 Of rotating shaft Cal 12-12 Stokes' law and density of circulation Cal 12-11 Total circulation and density of circulation Stokes' law Cal 12-11 Transients Cal 5-22 Particular solution Cal 5-22 Transient solution Cal 5-23 Traveling wave Sine wave Cal 2-14 Tritton, fluid dynamics text Cal 4-9 Turbulence Cal 7-10 Two dimensional conserved current Vortex, intuitive discussion of Cal 13 A2-2 Two dimensional vortex ring Cal 12-19 Speed of Cal 12-19

## U

Uncertainty principle Cal 1-2 And definition of velocity Cal 1-2 And strobe photographs Cal 1-2 Applied to projectile motion Cal 1-4 Uniquely determined field Conditions for Cal 12-2 Unit vectors Cylindrical coordinate system Cal 3-14 Derivative of changing unit vectors Cal 4-12 Spherical coordinate system Cal 3-16 Units CGS Centimeter, gram, second Back cover-1 MKS Meter, kilogram, second Front cover-2

## V

Variables Evaluated over interval Cal 1-10 Vector Components in rotated coordinate system Cal 3-26 Definition of acceleration Cal 1-7 Component equations Cal 1-8 Definition of velocity Cal 1-6 Component equations Cal 1-8 Dot product Summation convention, Einstein's Cal 13-5 Equations Components with derivatives Cal 1-7 Gradient vector Transformation of Cal 3-25 Vector fields As gradient of scalar field Cal 3-29 Created by gradient Cal 3-1 Gradient as a vector field Cal 3-28 Vector identities Cal 8-3 For a moving circuit Cal 13-12 Rate of change of flux through Cal 13-15 For use with Maxwell's equations Cal 9-2 Many of them Form.-4 Vector potential And the electric field Cal 11-3 Chapter on Cal 11-1 Curl and divergence of Cal 11-4 Divergent and solenoidal parts Cal 11-4 Gauge invarience Cal 11-4 In Faraday's law Cal 11-3 Introducing into Maxwell's equations Cal 11-3 Introduction to Cal 11-2 Needed in quantum theory Cal 11-3 Unneeded in classical electromagnetism Cal 11-3 Wave equation for Cal 11-4 Coulomb gauge Cal 11-6 Gauge invariant form Cal 11-5 Wave gauge Cal 11-5 Velocity And the uncertainty principle Cal 1-4 Calculus definition of Cal 1-5 Component equations Cal 1-8 Curve, area under Cal 1-12 Definite integral of Cal 1-11 Integral of Cal 1-10 Velocity field Derived from a potential Cal 12-3 Of a rotating shaft Solid body rotation Cal 12-9 Stokes' law applied to Cal 12-12

Viscosity Coefficient of Cal 4-4 Experimental formula for Cal 4-11 Measuring Cal 4-9 None in superfluids Cal 12-6 Role in Navier-Stokes equation Cal 13-7 Second viscosity coefficient Cal 4-6 Term in Navier-Stokes equation Cal 13-10 Viscous stress Cal 4-4 Viscous force Del squared Cal 4-1 In cylindrical coordinates Cal 4-7 In pipe flow Cal 4-7 Viscous force formula Cal 4-8 On a fluid element Cal 4-5 For 1D flows Cal 4-5 For 3D flows Cal 4-6 Where it acts in a vortex Cal 13-10 Voltage Electric field as gradient of Cal 3-4 Gradient of Equation for electric field Cal 3-7 Field of point charge Cal 3-10 Inside a conductor Cal 3-9 Interpretation Cal 3-6 Parallel plate capacitor Cal 3-8 Impedance formulas for Cal 5-18 In coaxial cable Cal 3-21 Voltage in circuit Driven LRC circuit Cal 5-19 Transient solutions Cal 5-22 In R.L.C circuits Differential Equations for Cal 5-6 Using impedance formulas Cal 5-16 Vortex Core, structure of Analogy to magnetic field Cal 12-10 Curl of velocity field Cal 12-9 Quantum Cal 12-15 Where viscous forces act Cal 13-10 Vortex currents Appendix on Cal 13 A2-1 Center of mass motion Cal 13 A2-5 Conserved, intuitive discussion of Cal 13 A2-2 Continuity equation for Cal 13 A2-3 Vortex dynamics Two ways to handle Cal 13 A2-1 Vortex dynamics equation Cal 13-12

Vortex line Center of mass motion Cal 13 A2-4 Gyroscope like behavior Cal 13-18 Magnus equation Airplane wing Cal 13-24 Relative motion of line and fluid Cal 13-20 Magnus formula for Cal 13 A2-7 Magnus lift force Cal 13-25 On fluid core vortices - a pseudo force Cal 13-26 Measurement of quantized circulation of Cal 13-20 Relative motion with fluid particles Cal 13-16 Magnus equation Cal 13-20 Sideways motion Caused by localized non potential force Cal 13-17 Single Cal 13 A2-4 Small unit flux tube of vorticity Cal 13-16 Vortex motion Effect of localized non potential force Cal 13-17 Vortex ring Creation of Cal 13 A2-9 Impulse of Cal 13-23 Impulse equation Cal 13-23 Ring does not carry linear momentum Cal 13-23 Motion of Conservation of energy in Cal 13-19 Motion of charged vortex ring Cal 13-18 Push forward to slow down Cal 13-19 Rayfield-Reif experiment Cal 13-16 Smoke ring Cal 12-20 Tubes of vorticity Cal 12-18 Two dimensional Cal 12-19 Vortex sheet Cal 12-13 Vortices Quantized, in superfluids Circulation of Cal 12-15 Core of Cal 12-16 Number in rotating bucket Cal 12-16 Vortex sheet Cal 12-13 Vorticity Analogous to magnetic field Cal 12-7 As a source of fluid motion Cal 12-7 Chapter on Cal 12-1 Conservation law for Cal 12-13 Continuity equation for Cal 13 A2-3 Creation of Cal 13 A2-9 Equation for Cal 13-12 In solid body rotation Cal 12-10 Vortex core structure Cal 12-10 Analogy to magnetic field Cal 12-10 Zero in potential flow Cal 12-3 Vorticity field Divergence of Cal 12-18 Flux in flow tube of vorticity Cal 12-18

## W

Wave Compressional wave on spring Cal 2-15 Pulse, formation of Cal 2-14 Rope Cal 2-10 Speed of waves Calculus derivation, rope Cal 2-13 One dimensional wave equation Cal 5-24 Spring compressional wave Cal 2-15 Wave equation Cal 2-1 Addition of waves Cal 2-14 Dirac's Cal 6-12 Electric field With sources Cal 11-6 For scalar and vector potentials Cal 11-4 For waves on a spring Cal 2-17 Introduction to Cal 2-10 Magnetic field With sources Cal 11-6 Maxwell's Derivation of Cal 9-4 Plane wave solution Cal 9-6 One dimensional Cal 2-1 General form of Cal 2-14 Solutions summary Cal 5-25 Solutions using complex variables Cal 5-24 Relativistic For zero rest mass particles Cal 6-2 Particles with rest mass Cal 6-3 Rope Cal 2-10 Scalar potential Coulomb gauge Cal 11-6 Gauge invariant form Cal 11-4 Wave gauge Cal 11-5 Schrödinger's Chapter on Cal 6-1 Discovery of Cal 6-1 Full three dimensional form Cal 6-6 Ideas that led to it Cal 6-2 Relativistic wave equation Cal 6-3 Sound Cal 2-17 The equation for Cal 2-20 Standing wave Cal 2-14 Three dimensional Cal 9-7 Vector potential Coulomb gauge Cal 11-6 Gauge invariant form Cal 11-5 Wave gauge Cal 11-5

Wave motion Sinusoidal, 1D Cal 5-24 Wave patterns Hydrogen Standing waves Cal 6-8 Wheel on fixed axle Cal 12-12 Stokes' law applied to Cal 12-13 Wing vortex Cal 13-24

## X

X.-Cal Ch 1 Exercise 1 Cal 1-14 Exercise 2 Cal 1-15 Exercise 3 Cal 1-17 Exercise 4 Cal 1-22 Exercise 5 Cal 1-24 Exercise 6 Cal 1-29 Exercise 7 Cal 1-29 Exercise 8 Cal 1-31 Exercise 9 Cal 1-33 Exercise 10 Cal 1-36 Exercise 11 Cal 1-39 Exercise 12 Cal 1-39 Exercise 13 Cal 1-39 X.-Cal Ch 2 Exercise 1 Cal 2-7 Exercise 2 Cal 2-8 Exercise 3 Cal 2-8 Exercise 4 Cal 2-9 X.-Cal Ch 3 Exercise 1 Cal 3-12 Exercise 2 Cal 3-15 Exercise 3 Cal 3-16 Exercise 4 Cal 3-21 X.-Cal Ch 3 view 2 Exercise 1 Cal 3-26 Exercise 2 Cal 3-27 X.-Cal Ch 4 Exercise 1 Cal 4-11 Exercise 2 Cal 4-17 Exercise 3 Cal 4-17 X.-Cal Ch 5 Exercise 1 Cal 5-5 Exercise 2 Cal 5-10 Exercise 3 Cal 5-12 Exercise 4 Cal 5-14

X.-Cal Ch 6 Exercise 1 Cal 6-3 Exercise 2 Cal 6-5 Exercise 3 Cal 6-7 Exercise 4 Cal 6-11 Exercise 5 Cal 6-15 Exercise 6 Cal 6-16 X.-Cal Ch 7 Exercise 1 Cal 7-4 Exercise 2 Cal 7-6 Exercise 3 Cal 7-9 Exercise 4 Cal 7-9 X.-Cal Ch 8 Exercise 1 Cal 8-5 Exercise 2 Cal 8-9 X.-Cal Ch 9 Exercise 1 Cal 9-8 X.-Cal Ch 11 Exercise 1 Cal 11-6 Exercise 2 Cal 11-6 X.-Cal Ch 12 Exercise 1 Cal 12-9 Exercise 2 Cal 12-10 Exercise 3 Cal 12-17 X.-Cal Ch 13 Exercise 1 Cal 13-18 Exercise 2 Cal 13-18 Exercise 3 Cal 13-19 Exercise 4 Cal 13-20

## Z

Zero rest mass particles Relativistic wave equation for Cal 6-2 Photons Cal 6-3

## Calculus Chapter 1 Introduction to Calculus

This first chapter covers all the calculus that is needed for the Physics 2000 text. The remaining chapters allow students to look at the physics from an advanced mathematical point of view.

This chapter, which replaces Chapter 4 in Physics 2000, is intended for students who have not had calculus, or as a calculus review for those whose calculus is not well remembered. If, after reading part way through this chapter, you feel your calculus background is not so bad after all, go back to Chapter 4 in Physics 2000, study the derivation of the constant acceleration formulas beginning on page 4-8, and work the projectile motion problems in the appendix to Chapter 4. Those who study all of this introduction to calculus should then proceed to the projectile motion problems in the appendix to Chapter 4 of the Physics text.

## LIMITING PROCESS

In Chapter 3 of Physics 2000, we used strobe photographs to define velocity and acceleration vectors. The basic approach was to turn up the strobe flashing rate, as we did in going from Figure (3-3) to (3-4) shown below. We turned the rate up until all the kinks are clearly visible and the successive displacement vectors give a reasonable description of the motion. We did not turn the flashing rate too high, for the practical reason that the displacement vectors became too short for accurate work.

In our discussion of instantaneous velocity we conceptually turned the strobe all the way up as illustrated in Figures (2-22a) through (2-22d), redrawn here in Figure (1). In these figures, we initially see a fairly large change in  $\vec{v}_0$  as the strobe rate is increased and  $\Delta t$  reduced. But then the change becomes smaller, and it looks as if we are approaching some final value of  $\vec{v}_0$  that does not depend on the size of  $\Delta t$ , provided  $\Delta t$  is small enough. It looks as if we have come close to the final value in Figure (1c).

The progression seen in Figure (1) is called a *limit-ing process*. The idea is that there really is some true value of  $\vec{v}_0$  which we have called the *instantaneous velocity*, and that we approach this true value for sufficiently small values of  $\Delta t$ . This is a calculus concept, and in the language of calculus, we are *taking the limit as*  $\Delta t$  goes to zero.

## THE UNCERTAINTY PRINCIPLE

For over 200 years, from the invention of calculus by Newton and Leibnitz until 1924, the limiting process and the resulting concept of instantaneous velocity was one of the cornerstones of physics. Then in 1924 Werner Heisenberg discovered what he called the *uncertainty principle* which places a limit on the accuracy of experimental measurements.

Heisenberg discovered something very new and unexpected. He found that the act of making an experimental measurement unavoidably affects the results of an experiment. This had not been known previously because the effect on large objects like golf balls is undetectable. But on an atomic scale where we study small systems like electrons moving inside an atom, the effect is not only observable, it can dominate our study of the system.

One particular consequence of the uncertainly principle is that the more accurately we measure the position of an object, the more we disturb the motion of the object. This has an immediate impact on the concept of instantaneous velocity. If we turn the strobe all the way up, reduce  $\Delta t$  to zero, we are in effect trying to measure the position of the object with infinite precision. The consequence would be an infinitely big disturbance of the motion of the object we are studying. If we actually could turn the strobe all the way up, we would destroy the object we were trying to study.



### Figures 3-3 and 3-4 from Physics 2000

Strobe photographs of a moving object. In the first photograph, the time between flashes is so long that the motion is difficult to understand. In the second, the time between flashes was reduced and the motion is more easily understood.



 $\Delta t = 0.025 \text{ Sec}$ 

## Figure 1

Transition to instantaneous velocity. As we reduce  $\Delta t$ , there is less and less change in the vector  $\vec{V}_0$ . It looks as if we are approaching an exact final value.

#### **Uncertainty Principle on a Larger Scale**

It turns out that the uncertainty principle can have a significant impact on a larger scale of distance than the atomic scale. Suppose, for example, we constructed a chamber that is 1 cm on each side, and wished to study the projectile motion of an electron inside. Using Galileo's idea that objects of different mass fall at the same rate, we would expect that the motion of the electron projectile should be the same as more massive objects. If we took a strobe photograph of the electron's motion, we would expect to get results like those shown in Figure (2). This figure represents projectile motion with an acceleration g = 980 cm/sec<sup>2</sup> and  $\Delta t = .01$ sec, as the reader can easily check.

When we study the uncertainty principle in Chapter 40 of the Physics text, we will see that a measurement which is accurate enough to show that position (2) is below position (1), could disturb the electron enough to reverse its direction of motion. The next position measurement could find the electron over where we drew position (3), or back where we drew position (0), or anywhere in the region in between. As a result we could not even determine what direction the electron is moving. This uncertainty would not be the result of a sloppy experiment, it is the best we can do with the most accurate and delicate measurements possible.

The uncertainty principle has had a significant impact on the way physicists think about motion. Because we now know that the measuring process affects the results of the measurement, we see that it is essential to provide experimental definitions to any physical quantity we wish to study. A conceptual definition, like turning the strobe all the way up to define instantaneous velocity, can lead to fundamental inconsistencies.

Even an experimental definition like our strobe definition of velocity can lead to inconsistent results when applied to something like the electron in Figure (2). But these inconsistencies are real. Their existence is telling us that the very concept of velocity is beginning to lose meaning for these small objects.

On the other hand, the idea of the limiting process and instantaneous velocity is very convenient when applied to larger objects where the effects of the uncertainty principle are not detectable. In this case we can apply all the mathematical tools of calculus developed over the past 250 years. The status of instantaneous velocity has changed from a basic concept to a useful mathematical tool. Those problems for which this mathematical tool works are called problems in *classical physics*; those problems for which the uncertainty principle is important, are in the realm of what we call *quantum physics*.



Figure 2

Hypothetical electron projectile motion experiment. The uncertainty principle tells us that such an experiment cannot lead to predictable results.

## CALCULUS DEFINITION OF VELOCITY

With the above perspective on the physical limitations of the limiting process, we can now return to the main topic of this chapter—the use of calculus in defining and working with velocity and acceleration.

In discussing the limiting process in calculus, one traditionally uses a special set of symbols which we can understand if we adopt the notation shown in Figure (3). In that figure we have drawn the coordinate vectors  $\vec{R}_i$  and  $\vec{R}_{i+1}$  for the i th and (i + 1) positions of the object. We are now using the symbol  $\Delta \vec{R}_i$  to represent the displacement of the ball during the i to i+1 interval. The vector equation for  $\Delta \vec{R}_i$  is

$$\vec{\Delta R}_i = \vec{R}_{i+1} - \vec{R}_i \tag{1}$$

In words, Equation (1) tells us that  $\overrightarrow{\Delta R}_i$  is the change, during the time  $\Delta t$ , of the position vector  $\vec{R}$  describing the location of the ball.



**Figure 3** Definitions of  $\overline{\Delta R}_i$  and  $\vec{v}_i$ .

The velocity vector  $\vec{v}_i$  is now given by

$$\vec{v}_i = \frac{\vec{\Delta R}_i}{\Delta t} \tag{2}$$

This is just our old strobe definition  $\vec{v}_i = \vec{s}_i/\Delta t$ , but using a notation which emphasizes that the displacement  $\vec{s}_i = \Delta \vec{R}_i$  is the *change in position* that occurs during the time  $\Delta t$ . The Greek letter  $\Delta$  (delta) is used both to represent the idea that the quantity  $\Delta \vec{R}_i$ or  $\Delta t$  is small, and to emphasize that both of these quantities change as we change the strobe rate.

The limiting process in Figure (1) can be written in the form

$$\underline{\vec{v}}_{i} \equiv \lim_{\Delta t \to 0} \frac{\overline{\Delta R}_{i}}{\Delta t}$$
(3)

where the word "limit" with  $\Delta t \rightarrow 0$  underneath, is to be read as "limit as  $\Delta t$  goes to zero". For example we would read Equation (3) as "*the instantaneous velocity*  $\vec{v}_i$  *at position* i *is the limit, as*  $\Delta t$  *goes to zero, of the ratio*  $\Delta \vec{R}_i / \Delta t$ ."

For two reasons, Equation (3) is not quite yet in standard calculus notation. One is that in calculus, only the limiting value, in this case, the instantaneous velocity, is considered to be important. Our strobe definition  $\vec{v}_i = \Delta \vec{R}_i / \Delta t$  is only a step in the limiting process. Therefore when we see the vector  $\vec{v}_i$ , we should assume that it is the limiting value, and no special symbol like the underline is used. For this reason we will drop the underline and write

$$\vec{v}_i = \lim_{\Delta t \to 0} \frac{\Delta \vec{R}_i}{\Delta t}$$
 (3a)
The second change deals with the fact that when  $\Delta t$  goes to zero we need an infinite number of time steps to get through our strobe photograph, and thus it is not possible to locate a position by counting time steps. Instead we measure the time t that has elapsed since the beginning of the photograph, and use that time to tell us where we are, as illustrated in Figure (4). Thus instead of using  $\vec{v}_i$  to represent the velocity at position i, we write  $\vec{v}(t)$  to represent the velocity at time t. Equation (3) now becomes

$$\vec{v}(t) = \lim_{\Delta t \to 0} \frac{\Delta \vec{R}(t)}{\Delta t}$$
 (3b)

where we also replaced  $\Delta \vec{R}_i$  by its value  $\Delta \vec{R}(t)$  at time t.

Although Equation (3b) is in more or less standard calculus notation, the notation is clumsy. It is a pain to keep writing the word "limit" with a  $\Delta t \rightarrow 0$  underneath. To streamline the notation, we replace the Greek letter  $\Delta$  with the English letter d as follows

$$\lim_{\Delta t \to 0} \frac{\overrightarrow{\Delta R}(t)}{\Delta t} \equiv \frac{d\vec{R}(t)}{dt}$$
(4)

(The symbol  $\equiv$  means *defined equal to*.) To a mathematician, the symbol  $d\vec{R}(t)/dt$  is just short-



#### Figure 4

Rather than counting individual images, we can locate a position by measuring the elapsed time t. In this figure, we have drawn the displacement vector  $\overline{R}(t)$  at time t = .3 sec. hand notation for the limiting process we have been describing. But to a physicist, there is a different, more practical meaning. Think of dt as a short  $\Delta t$ , short enough so that the limiting process has essentially occurred, but not too short to see what is going on. In Figure (1), a value of dt less than .025 seconds is probably good enough.

If dt is small but finite, then we know exactly what the  $d\vec{R}(t)$  is. It is the small but finite displacement vector at the time t. It is our old strobe definition of velocity, with the added condition that dt is such a short time interval that the limiting process has occurred. From this point of view, dt is a real time interval and  $d\vec{R}(t)$  a real vector, which we can work with in a normal way. The only thing special about these quantities is that when we see the letter d instead of  $\Delta$ , we must remember that a limiting process is involved. In this notation, the calculus definition of velocity is

$$\vec{v}(t) = \frac{d\vec{R}(t)}{dt}$$
(5)

where  $\dot{R}(t)$  and  $\vec{v}(t)$  are the particle's coordinate vector and velocity vector respectively, as shown in Figure (5). Remember that this is just fancy shorthand notation for the limiting process we have been describing.



**Figure 5** *Instantaneous position and velocity at time t.* 

# ACCELERATION

In the analysis of strobe photographs, we defined both a velocity vector  $\vec{v}$  and an acceleration vector  $\vec{a}$ . The definition of  $\vec{a}$ , shown in Figure (2-12) reproduced here in Figure (6), was

$$\vec{a}_{i} \equiv \frac{\vec{v}_{i+1} - \vec{v}_{i}}{\Delta t} \tag{6}$$

In our graphical work we replaced  $\vec{v}_i$  by  $\vec{s}_i/\Delta t$  so that we could work directly with the displacement vectors  $\vec{s}_i$  and experimentally determine the behavior of the acceleration vector for several kinds of motion.

Let us now change this graphical definition of acceleration over to a calculus definition, using the ideas just applied to the velocity vector. First, assume that the ball reached position i at time t as shown in Figure (6). Then we can write

$$\vec{v}_i = \vec{v}(t)$$
  
 $\vec{v}_{i+1} = \vec{v}(t+\Delta t)$ 

to change the time dependence from a count of strobe flashes to the continuous variable t. Next, define the vector  $\overrightarrow{\Delta v}(t)$  by

$$\vec{\Delta \mathbf{v}}(t) \equiv \vec{\mathbf{v}}(t + \Delta t) - \vec{\mathbf{v}}(t) \quad \left( = \vec{\mathbf{v}}_{i+1} - \vec{\mathbf{v}}_i \right)$$
(7)

We see that  $\overrightarrow{\Delta v}(t)$  is the change in the velocity vector as the time advances from t to  $t+\Delta t$ . The strobe definition of  $\vec{a}_i$  can now be written

$$\vec{a}(t) \begin{pmatrix} strobe\\ definition \end{pmatrix} = \frac{\vec{v}(t + \Delta t) - \vec{v}(t)}{\Delta t} \equiv \frac{\Delta \vec{v}(t)}{\Delta t} \quad (8)$$



**Figure 6** *Experimental definition of the acceleration vector.* 

Now go through the limiting process, turning the strobe up, reducing  $\Delta t$  until the value of  $\vec{a}(t)$  settles down to its limiting value. We have

$$\vec{a}(t) \begin{pmatrix} calculus\\ definition \end{pmatrix} = \lim_{\Delta t \to 0} \frac{\vec{v}(t + \Delta t) - \vec{v}(t)}{\Delta t}$$
$$= \lim_{\Delta t \to 0} \frac{\vec{\Delta v}(t)}{\Delta t}$$
(9)

Finally use the shorthand notation d/dt for the limiting process:

$$\vec{a}(t) = \frac{d\vec{v}(t)}{dt}$$
(10)

Equation (10) does not make sense unless you remember that it is notation for all the ideas expressed above. Again, physicists think of dt as a short but finite time interval, and  $d\vec{v}(t)$  as the small but finite change in the velocity vector during the time interval dt. It's our strobe definition of acceleration with the added requirement that  $\Delta t$  is short enough that the limiting process has already occurred.

## Components

Even if you have studied calculus, you may not recall encountering formulas for the derivatives of vectors, like  $d\vec{R}(t)/dt$  and  $d\vec{v}(t)/dt$  which appear in Equations (5) and (10). To bring these equations into a more familiar form where you can apply standard calculus formulas, we will break the vector Equations (5) and (10) down into component equations.

In the chapter on vectors, we saw that any vector equation like

$$\vec{A} = \vec{B} + \vec{C} \tag{11}$$

is equivalent to the three component equations

$$A_x = B_x + C_x$$
  

$$A_y = B_y + C_y$$
  

$$A_z = B_z + C_z$$
(12)

The advantage of the component equations was that they are simply numerical equations and no graphical work or trigonometry is required. The limiting process in calculus does not affect the decomposition of a vector into components, thus Equation (5) for  $\vec{v}(t)$  and Equation (10) for  $\vec{a}(t)$  become

$$\vec{v}(t) = d\vec{R}(t)/dt$$
 (5)

$$v_{x}(t) = dR_{x}(t)/dt$$
 (5a)

$$v_{y}(t) = dR_{y}(t)/dt$$
 (5b)

$$v_{z}(t) = dR_{z}(t)/dt$$
 (5c)

and

$$\vec{a}(t) = d\vec{v}(t)/dt \tag{10}$$

$$a_{x}(t) = dv_{x}(t)/dt$$
(10a)

$$a_v(t) = dv_v(t)/dt$$

$$a_{z}(t) = dv_{z}(t)/dt$$
 (10c)

(10b)

Often we use the letter x for the x coordinate of the vector  $\vec{R}$  and we use y for  $R_y$  and z for  $R_z$ . With this notation, Equation (5) assumes the shorter and perhaps more familiar form



At this point the notation has become deceptively short. You now have to remember that x(t) stands for the x coordinate of the particle at a time t.

We have finally boiled the notation down to the point where it would be familiar in any calculus course. If we restrict our attention to one dimensional motion along the x axis, then all we have to concern ourselves with are the x component equations

$$v_{x}(t) = \frac{dx(t)}{dt}$$

$$a_{x}(t) = \frac{dv_{x}(t)}{dt}$$
(10a)

## INTEGRATION

When we worked with strobe photographs, the photograph told us the position  $\vec{R}(t)$  of the ball as time passed. Knowing the position, we can then use Equation (5) to calculate the ball's velocity  $\vec{v}(t)$  and then Equation (10) to determine the acceleration  $\vec{a}(t)$ . In general, however, we want to go the other way, and predict the motion from a knowledge of the acceleration. For example, imagine that you were in Galileo's position, hired by a prince to predict the motion of cannonballs. You know that a cannonball should not be much affected by air resistance, thus the acceleration throughout its trajectory should be the constant gravitational acceleration g . You know that  $\vec{a}(t) = \vec{g}$ . How then do you use that knowledge in Equations (5) and (10) to predict the motion of the ball?

The answer is that you cannot with the equations in their present form. The equations tell you how to go from  $\vec{R}(t)$  to  $\vec{a}(t)$ , while to predict motion you need to go the other way, from  $\vec{a}(t)$  to  $\vec{R}(t)$ . The topic of this section is to see how to reverse the directions in which we use our calculus equations. Equations (5) and (10) involve the process called *differentiation*. We will see that when we go the other way the reverse of differentiation is a process called *integration*. We will see that integration is a simple concept, but a process that is sometimes hard to perform without the aid of a computer.

## **Prediction of Motion**

In our earlier discussion, we have used strobe photographs to analyze motion. Let us see what we can learn from such a photograph for predicting motion. Figure (8) is our familiar projectile motion photograph showing the displacement  $\vec{s}$  of a ball during the time the ball traveled from a position labeled (0) to the position labeled (4). If the ball is now at position (0) and each of the images is (.1) seconds apart, then the vector  $\vec{s}$  tells us where the ball will be at a time of (.4) seconds from now. If we can predict  $\vec{s}$ , we can predict the motion of the ball. The general problem of predicting the motion of the ball is to be able to calculate  $\vec{s}(t)$  for any time t.

From Figure (8) we see that  $\vec{s}$  is the vector sum of the individual displacement vectors  $\vec{s}_1$ ,  $\vec{s}_2$ ,  $\vec{s}_3$  and  $\vec{s}_4$ 

$$\vec{s} = \vec{s}_1 + \vec{s}_2 + \vec{s}_3 + \vec{s}_4 \tag{11}$$

We can then use the fact that  $\vec{s}_1=\vec{v}_1\Delta t$  ,  $\vec{s}_2=\vec{v}_2\Delta t$  , etc. to get

$$\vec{s} = \vec{v}_1 \Delta t + \vec{v}_2 \Delta t + \vec{v}_3 \Delta t + \vec{v}_4 \Delta t \qquad (12)$$

Rather than writing out each term, we can use the *summation sign*  $\Sigma$  to write

$$\vec{s} = \sum_{i=1}^{4} \vec{v}_i \Delta t \tag{12a}$$



**Figure 8** To predict the total displacement  $\vec{s}$ , we add up the individual displacements  $\vec{s}_i$ .

Equation (12) is approximate in that the  $\vec{v}_i$  are approximate (strobe) velocities, not the instantaneous velocities we want for a calculus discussion. In Figure (9) we improved the situation by cutting  $\Delta t$  to 1/4 of its previous value, giving us four times as many images and more accurate velocities  $\vec{v}_i$ .

We see that the displacement  $\vec{s}$  is now the sum of 16 vectors

$$\vec{s} = \vec{s}_1 + \vec{s}_2 + \vec{s}_3 + \dots + \vec{s}_{15} + \vec{s}_{16}$$
 (13)

Expressing this in terms of the velocity vectors  $\vec{v}_1$  to  $\vec{v}_{16}$  we have

$$\vec{s} = \vec{v}_1 \Delta t + \vec{v}_2 \Delta t + \vec{v}_3 \Delta t + ... + \vec{v}_{15} \Delta t + \vec{v}_{16} \Delta t$$
 (14)

or using our more compact notation

$$\vec{s} = \sum_{i=1}^{16} \vec{v}_i \Delta t \tag{14a}$$

While Equation (14) for  $\vec{s}$  looks quite different than Equation (12)—the sum of sixteen vectors instead of four—the displacement vectors  $\vec{s}$  in the two cases are exactly the same. Adding more intermediate images did not change where the ball was located at the time of t = .4 seconds. In going from Equation (12) to (14), what has changed as a result of shortening the time step  $\Delta t$ , is that *the individual velocity vectors*  $\vec{v}_i$  *become more nearly equal to the instantaneous velocity of the ball at each image*.



**Figure 9** With a shorter time interval, we add up more displacement vectors to get the total displacement  $\vec{s}$ .

If we reduced  $\Delta t$  again by another factor of 1/4, so that we had 64 images in the interval t=0 to t=.4 sec, the formula for  $\vec{s}$  would become

$$\vec{s} = \sum_{i=1}^{64} \vec{v}_i \Delta t \tag{15a}$$

where now the  $\vec{v}_i$  are still closer to representing the ball's instantaneous velocity. The more we reduce  $\Delta t$ , the more images we include, the closer each  $\vec{v}_i$ comes to the instantaneous velocity  $\vec{v}(t)$ . While adding more images gives us more vectors that we have to add up to get the total displacement  $\vec{s}$ , there is very little change in our formula for  $\vec{s}$ . If we had a million images, we would simply write

$$\vec{s} = \sum_{i=1}^{1000000} \vec{v}_i \Delta t$$
 (16a)

In this case the  $\vec{v}_i$  would be physically indistinguishable from the instantaneous velocity  $\vec{v}(t)$ . We have essentially reached a calculus limit, but we have problems with the notation. It is clearly inconvenient to label each  $\vec{v}_i$  and then count the images. Instead we would like notation that involves the instantaneous velocity  $\vec{v}(t)$  and expresses the beginning and end points in terms of the initial time  $t_i$  and final time  $t_f$ , rather than the initial and final image numbers i.

In the calculus notation, we replace the summation sign  $\Sigma$  by something that looks almost like the summation sign, namely the *integral sign*  $\int$ . (The French word for integration is the same as their word for summation.) Next we replaced the individual  $\vec{v}_i$  by the continuous variable  $\vec{v}(t)$  and finally express the end points by the initial time  $t_i$  and the final time  $t_f$ . The result is

$$\vec{s} = \sum_{i=1}^{n} \vec{v}_{i} \Delta t \rightarrow \begin{pmatrix} as \ the \ number \\ n \ becomes \\ infinitely \\ large \end{pmatrix} \int_{t_{i}}^{t_{f}} \vec{v}(t) dt \qquad (17)$$

Calculus notation is more easily handled, or is at least more familiar, if we break vector equations up into component equations. Assume that the ball started at position i which has components  $x_i = x(t_i)$  [read  $x(t_i)$  as "x at time  $t_i$ "] and  $y_i = y(t_i)$  as shown in Figure (10). The final position f is at  $x_f = x(t_f)$  and  $y_f = y(t_f)$ .

Thus the displacement  $\vec{s}$  has x and y components

$$s_x = x(t_f) - x(t_i)$$
  

$$s_y = y(t_f) - y(t_i)$$

Breaking Equation (17) into component equations gives

$$s_x = x(t_f) - x(t_i) = \int_{t_i}^{t_f} v_x(t) dt$$
 (18a)

$$s_y = y(t_f) - y(t_i) = \int_{t_i}^{t_f} v_y(t) dt$$
 (18b)

Here we will introduce one more piece of notation often used in calculus courses. On the left hand side of Equation (18a) we have  $x(t_f) - x(t_i)$  which we can think of as the variable x(t) evaluated over the interval of time from  $t_i$  to  $t_f$ . We will often deal with variables evaluated over some interval and have a special notation for that. We will write

$$x(t_{f}) - x(t_{i}) \equiv x(t) \Big|_{t_{i}}^{t_{f}}$$
(19)

You are to read the symbol  $x(t) |_{t_i}^{t_f}$  as "x of t evaluated from  $t_i$  to  $t_f$ ". We write the initial time  $t_i$  at the bottom of the vertical bar, the final time  $t_f$  at the top.



**Figure 10** Breaking the vector  $\vec{s}$  into components.

We use similar notation for any kind of variable, for example

$$f(\mathbf{x})\Big|_{\mathbf{x}_1}^{\mathbf{x}_2} \equiv f(\mathbf{x}_2) - f(\mathbf{x}_1)$$
 (19a)

(Remember to subtract when the variable is evaluated at the value at the bottom of the vertical bar.)

With this notation, our Equation (18) can be written

$$s_{x} = x(t) \Big|_{t_{i}}^{t_{f}} = \int_{t_{i}}^{t_{f}} v_{x}(t) dt$$
 (18a')

$$s_y = y(t) \Big|_{t_i}^{t_f} = \int_{t_i}^{t_f} v_y(t) dt \qquad (18b')$$

## **Calculating Integrals**

Equation (18) is nice and compact, but how do you use it? How do you calculate integrals? The key is to remember that an integral is just a fancy notation for a sum of terms, where we make the time step  $\Delta t$  very small. Keeping this in mind, we will see that there is a very easy way to interpret an integral.



Figure 11a

Strobe photograph of ball moving at constant velocity in x direction.



Figure 11b

Graph of  $v_x(t)$  versus t for the ball of Figure 11a.



**Figure 11c** Each  $v_x \Delta t$  is the area of a rectangle.

To get this interpretation, let us start with the simple case of a ball moving in a straight line, for instance, the x direction, at a constant velocity  $v_x$ . A strobe picture of this motion would look like that shown in Figure (11a).

Figure (11b) is a graph of the ball's velocity  $v_x(t)$  as a function of the time t. The vertical axis is the value of  $v_x$ , the horizontal axis is the time t. Since the ball is traveling at constant velocity,  $v_x$  has a constant value and is thus represented by a straight horizontal line. In order to calculate the distance that the ball has traveled during the time interval from  $t_i$  to  $t_f$ , we need to evaluate the integral

$$s_{x} = \int_{t_{i}}^{t_{f}} v_{x}(t) dt \qquad \begin{array}{c} \text{distance ball} \\ \text{travels in} \\ \text{time interval} \\ t_{i} \text{ to } t_{f} \end{array} \tag{18a}$$

To actually evaluate the integral, we will go back to our summation notation

$$s_{x} = \sum_{i_{\text{initial}}}^{i_{\text{final}}} v_{xi} \Delta t$$
(20)

and show individual time steps  $\Delta t$  in the graph of  $v_x$  versus t, as in Figure (11c).

We see that each term in Equation (20) is represented in Figure (11c) by a rectangle whose height is  $v_x$  and whose width is  $\Delta t$ . We have shaded in the rectangle representing the 7th term  $v_{x7}\Delta t$ . We see that  $v_{x7}\Delta t$  is just the *area* of the shaded rectangle, and it is clear that the sum of all the areas of the individual rectangles is the total area under the curve, starting at time  $t_i$  and ending at time  $t_f$ . Here we are beginning to see that the process of integration is equivalent to finding the area under a curve.

With a simple curve like the constant velocity  $v_x(t)$ in Figure (11c), we see by inspection that the total area from  $t_i$  to  $t_f$  is just the area of the complete rectangle of height  $v_x$  and width  $(t_f - t_i)$ . Thus

$$\mathbf{s}_{\mathbf{x}} = \mathbf{v}_{\mathbf{x}} \times (\mathbf{t}_{\mathbf{f}} - \mathbf{t}_{\mathbf{i}}) \tag{21}$$

This is the expected result for constant velocity, namely

distance  
traveled = velocity × time 
$$\int_{velocity}^{for} constant$$
 (21a)

To see that you are not restricted to the case of constant velocity, suppose you drove on a freeway due east (the x direction) starting at 9:00 AM and stopping for lunch at 12 noon. Every minute during your trip you wrote down the speedometer reading so that you had an accurate plot of  $v_x(t)$  for the entire morning, a plot like that shown in Figure (12). From such a plot, could you determine the distance  $s_x$  that you had traveled?

Your best answer is to multiply each value  $v_i$  of your velocity by the time  $\Delta t$  to calculate the average distance traveled each minute. Summing these up from the initial time  $t_i = 9:00$  AM to the final time  $t_f = noon$ , you have as your estimate

$$s_x \approx \sum_i v_{xi} \Delta t$$

(The symbol  $\approx$  means *approximately equal*.)

To get a more accurate value for the distance traveled, you should measure your velocity at shorter time intervals  $\Delta t$  and add up the larger number of smaller rectangles. The precise answer should be obtained in the limit as  $\Delta t$  goes to zero

$$s_{x} = \lim_{\Delta t \to 0} \sum_{i} v_{xi} \Delta t = \int_{t_{i}}^{t_{f}} v_{x}(t) dt$$
(22)

This limit is just the area under the curve that is supposed to represent the instantaneous velocity  $v_x(t)$ .



Figure 12

Plot of  $v_x(t)$  for a trip starting at 9:00 AM and finishing at noon. The distance traveled is the area under the curve.

Thus we can interpret the integral of a curve as the area under the curve even when the curve is not constant or flat. Mathematicians concern themselves with curves that are so wild that it is difficult or impossible to determine the area under them. Such curves seldom appear in physics problems.

While the basic idea of integration is simple—just finding the area under a curve—in practice it can be quite difficult to calculate the area. Much of an introductory calculus course is devoted to finding the formulas for the areas under various curves. There are also books called *tables of integrals* where you look up the formula for a curve and the table tells you the formula for the area under that curve.

In Chapter 16 of the Physics text, we will discuss a mathematical technique called *Fourier analysis*. This is a technique in which we can describe the shape of any continuous curve in terms of a sum of sine waves. (Why we want to do that will become clear then.) The process of Fourier analysis involves finding the area under some very complex curves, curves often involving experimental data for which we have no formula, only graphs. Such curves cannot be integrated by using a table of integrals, with the result that Fourier analysis was not widely used until the advent of the modern digital computer.

The computer made a difference, because we can find the area under almost any curve by breaking the curve into short pieces of length  $\Delta t$ , calculating the area  $v_i \Delta t$  of each narrow rectangle, and adding up the area of the rectangles to get the total area. If the curve is so wild that we have to break it into a million segments to get an accurate answer, that might be too hard to do by hand, but it usually a very simple and rapid job for a computer. Computers can be much more efficient than people at integration.

# The Process of Integrating

There is a language for the process of integration which we will now take you through. In each case we will check that the results are what we would expect from our summation definition, or the idea that an integral is the area under a curve.

The simplest integral we will encounter is the calculation of the area under a curve of unit height as shown in Figure (13). We have the area of a rectangle of height 1 and length  $(t_f - t_i)$ 



**Figure 13** *Area under a curve of unit height.* 

We will use some special language to describe this integration. We will say that the integral of dt is simply the time t, and that the integral of dt from  $t_i$  to  $t_f$  is equal to t evaluated from  $t_i$  to  $t_f$ . In symbols this is written as

$$\int_{t_{i}}^{t_{f}} dt = t \Big|_{t_{i}}^{t_{f}} = (t_{f} - t_{i})$$
(23)

Recall that the vertical line after a variable means to evaluate that variable at the final position  $t_f$  (upper value), minus that variable evaluated at the initial position  $t_i$  (lower value). Notice that this prescription gives the correct answer.

The next simplest integral is the integral of a constant, like a constant velocity  $v_x$  over the interval  $t_i$  to  $t_f$ 



**Figure 14** Area under the constant  $v_r$  curve.

Since  $(t_f - t_i) = \int_{t_i}^{t_f} dt$ , we can replace  $(t_f - t_i)$  in Equation (24) by the integral to get

$$\int_{t_i}^{t_f} v_x dt = v_x \int_{t_i}^{t_f} dt \qquad v_x a \ constant$$
(25)

and we see that a constant like  $v_x$  can be taken outside the integral sign.

Let us try the simplest case we can think of where  $v_x$  is not constant. Suppose  $v_x$  starts at zero at time  $t_i = 0$  and increases linearly according to the formula

$$v_x = at$$
 (26)





When we get up to the time  $t_f$  the velocity will be  $(at_f)$  as shown in Figure (15). The area under the curve  $v_x = at$  is a triangle whose base is of length  $t_f$  and height is  $at_f$ . The area of this triangle is one half the base times the height, thus we get for the distance  $s_x$  traveled by an object moving with this velocity

$$s_{x} = \int_{0}^{t_{f}} v_{x} dt = \frac{1}{2} (base) \times (height)$$
$$= \frac{1}{2} (t_{f}) (at_{f}) = \frac{1}{2} at_{f}^{2}$$
(27)

Now let us repeat the same calculation using the language one would find in a calculus book. We have

$$s_x = \int_0^{t_f} v_x dt = \int_0^{t_f} (at) dt$$
 (28)

The constant (a) can come outside, and we know that the answer is  $1/2at_f^2$ , thus we can write

$$s_x = a \int_0^{t_f} t dt = \frac{1}{2} a t_f^2$$
 (29)

In Equation (29) we can cancel the a's to get the result

$$\int_{0}^{t_{x}} t dt = \frac{1}{2} t_{f}^{2}$$
(30)

In a calculus text, you would find the statement that the integral  $\int t dt$  is equal to  $t^2/2$  and that the integral should be evaluated as follows

$$\int_{0}^{t_{\rm f}} t dt = \frac{t^2}{2} \bigg|_{0}^{t_{\rm f}} = \frac{t_{\rm f}^2}{2} - \frac{0}{2} = \frac{t_{\rm f}^2}{2}$$
(31)

# **Indefinite Integrals**

When we want to measure an actual area under a curve, we have to know where to start and stop. When we put these limits on the integral sign, like  $t_i$  and  $t_f$ , we have what is called a *definite integral*. However there are times where we just want to know what the form of the integral is, with the idea that we will put in the limits later. In this case we have what is called an *indefinite integral*, such as

$$\int t dt = \frac{t^2}{2} \quad indefinite integral \tag{32}$$

The difference between our definite integral in Equation (31) and the indefinite one in Equation (32) is that we have not chosen the limits yet in Equation (32). If possible, a table of integrals will give you a formula for the indefinite integral and let you put in whatever limits you want.

# **Integration Formulas**

For some sets of curves, there are simple formulas for the area under them. One example is the set of curves of the form  $t^n$ . We have already considered the cases where n = 0 and n = 1.

$$n = 0$$

$$\int t^{0} dt = \int dt = t$$
(33a)

$$n = I$$

$$\int t^{1} dt = \int t dt = \frac{t^{2}}{2}$$
(33b)

Some results we will prove later are

$$n = 2$$

$$\int t^2 dt = \frac{t^3}{3}$$

$$t^2 \qquad (33c)$$

$$n = 3$$

$$\int t^{3} dt = \frac{t^{4}}{4}$$
(33d)

Looking at the way these integrals are turning out, we suspect that the general rule is

$$\int t^n dt = \frac{t^{n+1}}{n+1}$$
(34)

It turns out that Equation (34) is a general result for any value of n except n = -1. If n = -1, then you would have division by zero, which cannot be the answer. (We will shortly discuss the special case where n = -1.)

As long as we stay away from the n = -1 case, the formula works for negative numbers. For example

$$\int t^{-2} dt = \int \frac{dt}{t^2} = \frac{t^{(-2+1)}}{-2+1} = \frac{t^{-1}}{(-1)}$$

$$\int \frac{dt}{t^2} = -\frac{1}{t}$$
(35)

In our discussion of gravitational and electrical potential energy, we will encounter integrals of the form seen in Equation (35).

#### **Exercise 1**

Using Equation (34) and the fact that constants can come outside the integral, evaluate the following integrals:

(a) 
$$\int x dx$$
 it does not matter whether  
we call the variable t or x  
(b)  $\int_{x=1}^{x=2} x^5 dx$  also sketch the area  
being evaluated  
(c)  $\int_{t=1}^{t=2} \frac{dt}{t^2}$  show that you get  
a positive area  
(d)  $\int \frac{GmM}{r^2} dr$  where G, m, and M  
are constants  
(e)  $\int \frac{a}{y^{3/2}} dy$  "a" is a constant

# **NEW FUNCTIONS**

We have seen that when we integrate a curve or function like  $t^2$ , we get a new function  $t^3/3$ . The functions  $t^2$  and  $t^3$  appear to be fairly similar; the integration did not create something radically different. However, the process of integration can lead to some curves with entirely different behavior. This happens, for example, in that special case n = -1 when we try to do the integral of  $t^{-1}$ .

# Logarithms

It is certainly not hard to plot  $t^{-1}$ , the result is shown in Figure (16). Also there is nothing fundamentally difficult or peculiar about measuring the area under the  $t^{-1}$  curve from some  $t_i$  to  $t_f$ , as long as we stay away from the origin t = 0 where  $t^{-1}$  blows up. The formula for this area turns out, however, to be the new function called the *natural logarithm*, abbreviated by the symbol *ln*. The area in Figure (16) is given by the formula

$$\int_{t_{i}}^{t_{f}} \frac{1}{t} dt = \ln(t_{f}) - \ln(t_{i})$$
(36)



Plot of  $t^{-1}$ . The area under this curve is the natural logarithm ln.

Two of the important but peculiar features of the natural logarithm are

$$\ln(ab) = \ln(a) + \ln(b) \tag{37}$$

$$\ln(\frac{1}{a}) = -\ln(a) \tag{38}$$

Thus we get, for example

$$\ln(t_{f}) - \ln(t_{i}) = \ln(t_{f}) + \ln\left(\frac{1}{t_{i}}\right)$$
$$= \ln\left(\frac{t_{f}}{t_{i}}\right)$$
(39)

Thus the area under the curve in Figure (16) is

$$\int_{t_i}^{t_f} \frac{dt}{t} = \ln\left(\frac{t_f}{t_i}\right)$$
(40)

While the natural logarithm has some rather peculiar properties it is easy to evaluate because it is available on all scientific calculators. For example, if  $t_i = .5$  seconds and  $t_f = 4$  seconds, then we have

$$\ln\left(\frac{t_{f}}{t_{i}}\right) = \ln\left(\frac{4}{.5}\right) = \ln\left(8\right)$$
(41)

Entering the number 8 on a scientific calculator and pressing the button labeled *ln*, gives

$$\ln(8) = 2.079 \tag{42}$$

which is the answer.

### Exercise 2

Evaluate the integrals

$$\int_{.001}^{1000} \frac{dx}{x} \qquad \int_{.000001}^{1} \frac{dx}{x}$$

Why are the answers the same?

# **The Exponential Function**

We have just seen that, while the logarithm function may have some peculiar properties, it is easy to evaluate using a scientific calculator. The question we now want to consider is whether there is some function that undoes the logarithm. When we enter the number 8 into the calculator and press ln, we get the number 2.079. Now we are asking if, when we enter the number 2.079, can we press some key and get back the number 8? The answer is, you press the key labeled  $e^x$ . The  $e^x$  key performs the *exponential function* which undoes the logarithm function. We say that the exponential function  $e^x$  is the *inverse* of the logarithm function ln.

# **Exponents to the Base 10**

You are already familiar with exponents to the base 10, as in the following examples

$$10^{0} = 1$$

$$10^{1} = 10$$

$$10^{-1} = 1/10 = .1$$

$$10^{2} = 100$$

$$10^{-2} = 1/100 = .01$$

$$10^{6} = 1,000,000$$

$$10^{-6} = .000001$$
(43)

The exponent, the number written above the 10, tells us how many factors of 10 are involved. A minus sign means how many factors of 10 we divide by. From this alone we deduce the following rules for the exponent to the base 10.

$$10^{-a} = \frac{1}{10^{a}} \tag{44}$$

$$10^{a} \times 10^{b} = 10^{a+b} \tag{45}$$

(Example  $10^2 \times 10^3 = 100 \times 1000 = 100,000$ .)

The inverse of the *exponent to the base 10* is the function called *logarithm to the base 10* which is denoted by the key labeled *log* on a scientific calculator. Formally this means that

$$\log\left(10^{\mathrm{y}}\right) = \mathrm{y} \tag{46}$$

Check this out on your scientific calculator. For example, enter the number 1,000,000 and press the *log* button and see if you get the number 6. Try several examples so that you are confident of the result.

# The Exponential Function y<sup>x</sup>

Another key on your scientific calculator is labeled  $y^x$ . This allows you to determine the value of any number y raised to the power (or exponent) x. For example, enter the number y = 10, and press the  $y^x$  key. Then enter the number x = 6 and press the = key. You should see the answer

$$y^{x} = 10^{6} = 1000000$$

It is quite clear that all exponents obey the same rules we saw for powers of 10, namely

$$y^a \times y^b = y^{a+b} \tag{47}$$

[Example  $y^2 \times y^3 = (y \times y)(y \times y \times y) = y^5$ .]

And as before

$$y^{-a} \equiv \frac{1}{y^a} \tag{48}$$

#### Exercise 3

Use your scientific calculator to evaluate the following quantities. (You should get the answers shown.)

- (a)  $10^6$  (100000) (b)  $2^3$  (8) (c)  $23^0$  (1)
- (d)  $10^{-1}$  (.1)

(To do this calculation, enter 10, then press  $y^x$ . Then enter 1, then press the +/- key to change it to -1, then press = to get the answer .1)

(e) 2 <sup>5</sup>	(1/√2=.707)
(f) log(10)	(1)
(g) ln (2.7183)	(1) (very close to 1)

Try some other examples on your own to become completely familiar with the  $y^x$  key. (You should note that any positive number raised to the 0 power is 1. Also, some calculators, in particular the one I am using, cannot handle any negative values of y, not even  $(-2)^2$  which is +4)

# Euler's Number e = 2.7183...

We have seen that the function *log* on the scientific calculator undoes, is the inverse of, powers of 10. For example, we saw that

$$\log (10^{x}) = x$$
 (46) repeated

Example:  $\log(10^6) = 6$ 

Earlier we saw that the exponential function  $e^x$  was the inverse of the natural logarithm *ln*. This means that

$$\ln(e^{x}) = x \tag{49}$$

The difference between the logarithm log and the natural logarithm ln, is that log undoes exponents of the number 10, while ln undoes exponents of the number e. This special number e, one of the fundamental mathematical constants like  $\pi$ , is known as **Euler's number**, and is always denoted by the letter e.

You can find the numerical value of Euler's number *e* on your calculator by evaluating

$$e^1 = e \tag{50}$$

To do this, enter 1 into your calculator, press the e<sup>x</sup> key, and you should see the result

$$e^1 = e = 2.718281828$$
 (51)

We will run into this number throughout the course. You should remember that *e* is about 2.7, or you might even remember 2.718. (Only remembering *e* as 2.7 is as klutzy as remembering  $\pi$  as 3.1)

The terminology in math courses is that the function *log*, which undoes exponents of the number 10, is the *logarithm to the base 10*. The function *ln*, what we have called the *natural logarithm*, which undoes exponents of the number *e*, is the *logarithm to the base e*. You can have logarithms to any base you want, but in practice we only use base 10 (because we have 10 fingers) and the base *e*. The base *e* is special, in part because that is the logarithm that naturally arises when we integrate the function 1/x. We will see shortly that the functions *ln* and  $e^x$  have several more, very special features.

## DIFFERENTIATION AND INTEGRATION

The scientific calculator is a good tool for seeing how the functions like ln and  $e^x$  are inverse of each other. Another example of inverse operations is integration and differentiation. We have seen that integration allows us to go the other way from differentiation [finding x(t) from v(t), rather than v(t) from x(t)]. However it is not so obvious that integration and differentiation are inverse operations when you think of integration as finding the area under a curve, and differentiation as finding limits of  $\Delta x/\Delta t$  as  $\Delta t$  goes to zero. It is time now to make this relationship clear.

First, let us review our concept of a derivative. Going back to our strobe photograph of Figure (3), replacing  $\vec{R}_i$  by  $\vec{R}(t)$  and  $\vec{R}_{i+1}$  by  $\vec{R}(t+\Delta t)$ , as shown in Figure (3a), our strobe velocity was then given by

$$\vec{v}(t) = \frac{\vec{R}(t+\Delta t) - \vec{R}(t)}{\Delta t}$$
 (52)

The calculus definition of the velocity is obtained by reducing the strobe time interval  $\Delta t$  until we obtain the instantaneous velocity  $\vec{v}$ .

$$\vec{v}_{calculus} = \lim_{\Delta t \to 0} \frac{\vec{R}(t + \Delta t) - \vec{R}(t)}{\Delta t}$$
 (53)



**Figure 3a** Defining the strobe velocity.

While Equation (53) looks like it is applied to the explicit case of the strobe photograph of projectile motion, it is easily extended to cover any process of differentiation. Whatever function we have [we had  $\vec{R}(t)$ , suppose it is now f(t)], evaluate it at two closely spaced times, subtract the older value from the newer one, and divide by the time difference  $\Delta t$ . Taking the limit as  $\Delta t$  becomes very small gives us the derivative

$$\frac{\mathrm{d}f(t)}{\mathrm{d}t} \equiv \lim_{\Delta t \to 0} \frac{f(t + \Delta t) - f(t)}{\Delta t}$$
(54)

The variable with which we are differentiating does not have to be time t. It can be any variable that we can divide into small segments, such as x

$$\frac{d}{dx}f(x) \equiv \lim_{\Delta x \to 0} \frac{f(x + \Delta x) - f(x)}{\Delta x}$$
(55)

Let us see how the operation defined in Equation (55) is the inverse of finding the area under a curve.

Suppose we have a curve, like our old  $v_x(t)$  graphed as a function of time, as shown in Figure (17). To find out how far we traveled in a time interval from  $t_i$  to some later time T, we would do the integral

$$\mathbf{x}(\mathbf{T}) = \int_{t_i}^{\mathbf{T}} \mathbf{v}_{\mathbf{x}}(t) \, \mathrm{d}t \tag{56}$$

The integral in Equation (56) tells us how far we have gone at any time T during the trip. The quantity x(T) is a function of this time T.



Figure 17

The distance traveled by the time T is the area under the velocity curve up to the time T.

Now let us differentiate the function x(T) with respect to the variable T. By our definition of differentiation we have

$$\frac{d}{dT}x(T) = \lim_{\Delta t \to 0} \frac{x(T + \Delta t) - x(T)}{\Delta t}$$
(57)

Figure (17) shows us the function x(T). It is the area under the curve v(t) starting at  $t_i$  and going up to time t = T. Figure (18) shows us the function  $x(T + \Delta t)$ . It is the area under the same curve, starting at  $t_i$  but going up to  $t = T + \Delta t$ . When we subtract these two areas, all we have left is the area of the slender rectangle shown in Figure (19).



**Figure 17 repeated** *The distance* x(T) *traveled by the time* T



Figure 18

The distance  $x(T+\Delta t)$  traveled by the time  $T+\Delta t$ .



The distance  $x(T+\Delta t) - x(T)$ traveled during the time  $\Delta t$ .

The rectangle has a height approximately  $v_x(T)$  and a width  $\Delta t$  for an area

$$x(T + \Delta t) - x(T) = v_x(T)\Delta t$$
(58)

Dividing through by  $\Delta t$  gives

$$v_{x}(T) = \frac{x(T + \Delta t) - x(T)}{\Delta t}$$
(59)

The only approximation in Equation (59) is at the top of the rectangle. If the curve is not flat,  $v_x(T + \Delta t)$  will be different from  $v_x(T)$  and the area of the sliver will have a value somewhere between  $v_x(T)\Delta t$  and  $v_x(T + \Delta t)\Delta t$ . But if we take the limit as  $\Delta t$  goes to zero, the value of  $v_x(T + \Delta t)$  must approach  $v_x(T)$ , and we end up with the exact result

$$v_{x}(T) = \lim_{\Delta t \to 0} \frac{x(T + \Delta t) - x(T)}{\Delta t}$$
 (60)

This is just the derivative dx(t)/dt evaluated at t = T.

$$\left| \mathbf{v}_{\mathbf{x}}(\mathbf{T}) = \frac{\mathbf{d}\mathbf{x}(\mathbf{t})}{\mathbf{d}\mathbf{t}} \right|_{\mathbf{t}} = \mathbf{T}$$
(61a)

where we started from

$$x(T) = \int_{t_i}^{T} v_x(t) dt$$
(61b)

Equations (61a) and (61b) demonstrate explicitly how differentiation and integration are inverse operations. The derivative allowed us to go from x(t) to  $v_x(t)$  while the integral took us from  $v_x(t)$  to x(t). This inverse is not as simple as pushing a button on a calculator to go from *ln* to  $e^x$ . Here we have to deal with limits on the integration and a shift of variables from t to T. But these two processes do allow us to go back and forth.

#### A Fast Way to go Back and Forth

We introduced our discussion of integration by pointing out that equations

$$v_x(t) = \frac{dx(t)}{dt};$$
  $a_x(t) = \frac{dv_x(t)}{dt}$  (62a,b)

went the wrong way in that we were more likely to know the acceleration  $a_x(t)$  and from that want to calculate the velocity  $v_x(t)$  and distance traveled x(t). After many steps, we found that integration was what we needed.

We do not want to repeat all those steps. Instead we would like a quick and simple way to go the other way around. Here is how you do it. Think of the dt in (62a) as a small but finite time interval. That means you can treat it like any other number and multiply both sides of Equation (62a) through by it.

$$v_{x}(t) = \frac{dx(t)}{dt}$$
$$dx(t) = v_{x}(t)dt$$
(63)

Now integrate both sides of Equation (63) from some initial time  $t_i$  to a final time T. (If you do the same thing to both sides of an equation, both sides should still be equal to each other.)

$$\int_{t_i}^{T} dx(t) = \int_{t_i}^{T} v_x(t) dt$$
(64)

If dt is to be thought of as a small but finite time step, then dx(t) is the small but finite distance we moved in the time dt. The integral on the left side of Equation (64) is just the sum of all these short distances moved, which is just the total distance moved during the time from  $t_i$  to T.

$$\int_{t_i}^{T} dx(t) = x(t) \Big|_{t_i}^{T} = x(T) - x(t_i)$$
(65)

Thus we end up with the result

$$\mathbf{x}(t)\Big|_{t_{i}}^{\mathrm{T}} = \int_{t_{i}}^{\mathrm{T}} \mathbf{v}_{\mathbf{x}}(t)dt$$
(66)

Equation (66) is a little more general than (62b) for it allows for the fact that  $x(t_i)$  might not be zero. If,

however, we say that we started our trip at  $x(t_i) = 0$ , then we get the result

$$x(T) = \int_{t_i}^{T} v_x(t) dt$$
(67)

representing the distance traveled since the start of the trip.

# **Constant Acceleration Formulas**

The constant acceleration formulas, so well known from high school physics courses, are an excellent application of the procedures we have just described.

We will begin with motion in one dimension. Suppose a car is traveling due east, in the x direction, and for a while has a constant acceleration  $a_x$ . The car passes us at a time  $t_i = 0$ , traveling at a speed  $v_{x0}$ . At some later time T, if the acceleration  $a_x$  remains constant, how far away from us will the car be?

We start with the equation

$$a_{x}(t) = \frac{dv_{x}(t)}{dt}$$
(68)

Multiplying through by dt to get

$$dv_x(t) = a_x(t)dt$$

then integrating from time  $t_i = 0$  to time  $t_f = T$ , we get

$$\int_0^T dv_x(t) = \int_0^T a_x(t)dt$$
(69)

Since the integral  $\int dv_x(t) = v_x(t)$ , we have

$$\int_{0}^{T} dv_{x}(t) = v_{x}(t) \Big|_{0}^{T} = v_{x}(T) - v_{x}(0)$$
 (70)

where  $v_x(0)$  is the velocity  $v_{x0}$  of the car when it passed us at time t = 0.

While we can always do the left hand integral in Equation (69), we cannot do the right hand integral until we know  $a_x(t)$ . For the constant acceleration problem, however, we know that  $a_x(t) = a_x$  is constant, and we have

$$\int_0^T a_x(t)dt = \int_0^T a_x dt$$
(71)

Since constants can come outside the integral sign, we get

$$\int_{0}^{T} a_{x} dt = a_{x} \int_{0}^{T} dt = a_{x} t \Big|_{0}^{T} = a_{x} T$$
 (72)

where we used  $\int dt = t$ . Substituting Equations (70) and (72) in (69) gives

$$\mathbf{v}_{\mathbf{x}}\mathbf{T} - \mathbf{v}_{\mathbf{x}\mathbf{0}} = \mathbf{a}_{\mathbf{x}}\mathbf{T} \tag{73}$$

Since Equation (73) applies for any time T, we can replace T by t to get the well known result

$$v_x(t) = v_{x0} + a_x t$$
 (a<sub>x</sub> constant) (74)

Equation (74) tells us the speed of the car at any time t after it passed us, as long as the acceleration remains constant.

To find out how far away the car is, we start with the equation

$$v_{x}(t) = \frac{dx(t)}{dt}$$
(62a)

Multiplying through by dt to get

 $d\mathbf{x}(t) = \mathbf{v}_{\mathbf{x}}(t) dt$ 

then integrating from time t = 0 to time t = T gives (as we saw earlier)

$$\int_{0}^{T} dx(t) = \int_{0}^{T} v_{x}(t) dt$$
 (75)

The left hand side is

$$\int_{0}^{T} dx(t) = x(t) \Big|_{0}^{T} = x(T) - x(0)$$
(76)

If we measure along the x axis, starting from where we are (where the car was at t = 0) then x(0) = 0.

In order to do the right hand integral in Equation (75), we have to know what the function  $v_x(t)$  is. But for constant acceleration, we have from Equation (74)  $v_x(t) = v_{x0} + a_x t$ , thus

$$\int_{0}^{T} v_{x}(t) dt = \int_{0}^{T} (v_{x0} + a_{x}t) dt$$
 (77)

One of the results of integration that you should prove for yourself (just sketch the areas) is the rule

$$\int_{i}^{f} \left[ a(x) + b(x) \right] dx = \int_{i}^{f} a(x) dx + \int_{i}^{f} b(x) dx \quad (78)$$

thus we get

$$\int_{0}^{T} (v_{x0} + a_x t) dt = \int_{0}^{T} v_{x0} dt + \int_{0}^{T} a_x t dt \quad (79)$$

Since constants can come outside the integrals, this is equal to

$$\int_{0}^{T} (v_{x0} + a_{x}t)dt = v_{x0} \int_{0}^{T} dt + a_{x} \int_{0}^{T} t \, dt \quad (80)$$

Earlier we saw that

$$\int_{0}^{T} dt = t \Big|_{0}^{T} = T - 0 = T$$
 (23)

$$\int_{0}^{T} t dt = \frac{t^{2}}{2} \bigg|_{0}^{T} = \frac{T^{2}}{2} - 0 = \frac{T^{2}}{2}$$
(30)

Thus we get

$$\int_{0}^{T} (v_{x0} + a_x t) dt = v_{x0} T + \frac{1}{2} a_x T^2$$
(81)

Using Equations (76) and (81) in (75) gives

$$x(T) - x_0 = v_{x0}T + \frac{1}{2}a_xT^2$$

Taking  $x_0 = 0$  and replacing T by t gives the other constant acceleration formula

$$\mathbf{x}(t) = \mathbf{v}_{\mathbf{x}0}t + \frac{1}{2}\mathbf{a}_{\mathbf{x}}t^2 \qquad (\mathbf{a}_{\mathbf{x}} \textit{ constant}) \qquad (82)$$

You can now see that the factor of  $t^2/2$  in the constant acceleration formulas comes from the integral  $\int t dt$ .

#### **Exercise 4**

Find the formula for the velocity v(t) and position x(t) for a car moving with constant acceleration  $a_x$ , that was located at position  $x_i$  at some initial time  $t_i$ .

Start your calculation from the equations

$$v_{x}(t) = \frac{dx(t)}{dt}$$
$$a_{x}(t) = \frac{dv_{x}(t)}{dt}$$

and go through all the steps that we did to get Equations (74) and (82). See if you can do this without looking at the text.

If you have to look back to see what some steps are, then finish the derivation looking at the text. Then a day or so later, clean off your desk, get out a blank sheet of paper, write down this problem, put the book away and do the derivation. Keep doing this until you can do the derivation of the constant acceleration formulas without looking at the text.

# **Constant Acceleration Formulas in Three Dimensions**

dt

To handle the case of motion with constant acceleration in three dimensions, you start with the separate equations

$$v_{x}(t) = \frac{dx(t)}{dt} \qquad a_{x}(t) = \frac{dv_{x}(t)}{dt}$$
$$v_{y}(t) = \frac{dy(t)}{dt} \qquad a_{y}(t) = \frac{dv_{y}(t)}{dt} \qquad (83)$$
$$v_{z}(t) = \frac{dz(t)}{dt} \qquad a_{z}(t) = \frac{dv_{z}(t)}{dt}$$

dt

$$\begin{aligned} x(t) &= v_{x0}t + \frac{1}{2}a_xt^2 \qquad v_x(t) = v_{x0} + a_xt \\ y(t) &= v_{y0}t + \frac{1}{2}a_yt^2 \qquad v_y(t) = v_{y0} + a_yt \quad (84) \\ z(t) &= v_{z0}t + \frac{1}{2}a_zt^2 \qquad v_z(t) = v_{z0} + a_zt \end{aligned}$$

The final step is to combine these six equations into the two vector equations

$$\vec{x}(t) = \vec{v}_0 t + \frac{1}{2} \vec{a} t^2$$
;  $\vec{v}(t) = \vec{v}_0 + \vec{a} t$  (85)

These are the equations we analyzed graphically in Chapter 3 of the Physics text, in Figure (3-34) and Exercise (3-9). (There we wrote  $\vec{s}$  instead of  $\vec{x}(t)$ , and  $\vec{v}_i$  rather than  $v_0$ .)

In many introductory physics courses, considerable emphasis is placed on solving constant acceleration problems. You can spend weeks practicing on solving these problems, and become very good at it. However, when you have done this, you have not learned very much physics because most forms of motion are not with constant acceleration, and thus the formulas do not apply. The formulas were important historically, for they were the first to allow the accurate prediction of motion (of cannonballs). But if too much emphasis is placed on these problems, students tend to use them where they do not apply. For this reason we have placed the exercises using the constant acceleration equations in an appendix at the end of Chapter 4 of the Physics text. There are plenty of problems there for all the practice you will need with these equations. Doing these exercises requires only algebra, there is no practice with calculus. To get some experience with calculus, be sure that you can confidently do Exercise 4.

# MORE ON DIFFERENTIATION

In our discussion of integration, we saw that the basic idea was that the integral of some curve or function f(t) was equal to the area under that curve. That is an easy enough concept. The problems arose when we actually tried to find the formulas for the areas under various curves. The only areas we actually calculated were the rectangular area under f(t) = constant and the triangular area under f(t) = at. It was perhaps a surprise that the area under the simple curve 1/t should turn out to be a logarithm.

For differentiation, the basic idea of the process is given by the formula

$$\frac{df(t)}{dt} = \lim_{\Delta t \to 0} \frac{f(t + \Delta t) - f(t)}{\Delta t} \quad (54) \text{ repeated}$$

Equation (54) is short hand notation for a whole series of steps which we introduced through the use of strobe photographs. The basic idea of differentiation is more complex than integration, but, as we will now see, it is often a lot easier to find the derivative of a curve than its integral.

## Series Expansions

An easy way to find the formula for the derivative of a curve is to use a series expansion. We will illustrate the process by using the binomial expansion to calculate the derivative of the function  $x^n$ where n is any constant.

We used the binomial expansion, or at least the first two terms, in Chapter 1 of the Physics text. That was during our discussion of the approximation formulas that are useful in relativistic calculations. As we mentioned in Exercise (1-5), the binomial expansion is

$$(x + \alpha)^{n} = x^{n} + n\alpha x^{n-1} + \frac{n(n-1)}{2!} \alpha^{2} x^{n-2} \cdots$$
(86)

When  $\alpha$  is a number much smaller than 1 ( $\alpha < < 1$ ), we can neglect  $\alpha^2$  compared to  $\alpha$  (if  $\alpha = .01$ ,  $\alpha^2 = .0001$ ), with the result that we can accurately approximate  $(x + \alpha)^n$  by

$$(x + \alpha)^n \approx x^n + n\alpha x^{n-1}$$
  $\alpha \ll 1$  (87)

Equation (87) gives us all the approximation formulas found in Equations (1-20) through (1-25) on page 1-28 of the Physics text.

As an example of Equation (87), just to see that it works, let us take x = 5, n = 7 and  $\alpha = .01$  to calculate  $(5.01)^7$ . From the calculator we get

$$(5.01)^7 = 79225.3344 \tag{88}$$

(To do this enter 5.01, press the  $y^x$  button, then enter 7 and press the = button.) Let us now see how this result compares with

$$(x + \alpha)^n \approx x^n + n\alpha x^{n-1}$$
  
(5 + .01)<sup>7</sup>  $\approx 5^7 + 7(.01)5^6$  (89)

We have

$$5^7 = 78125$$
 (90)

$$7 \times .01 \times 5^{6} = 7 \times .01 \times 15625 = 1093.75 \quad (91)$$

Adding the numbers in (90) and (91) together gives

$$5^7 + 7(.01)5^6 = 79218.75$$
 (92)

Thus we end up with 79218 instead of 79225, which is not too bad a result. The smaller  $\alpha$  is compared to one, the better the approximation.

# Derivative of the Function x<sup>n</sup>

We are now ready to use our approximation formula (87) to calculate the derivative of the function  $x^n$ . From the definition of the derivative we have

$$\frac{d(x^{n})}{dx} = \lim_{\Delta x \to 0} \frac{(x + \Delta x)^{n} - x^{n}}{\Delta x}$$
(93)

Since  $\Delta x$  is to become infinitesimally small, we can use our approximation formula for  $(x + \alpha)^n$ . We get

$$(x + \alpha)^{n} \approx x^{n} + n(\alpha)x^{n-1} \qquad (\alpha << 1)$$
$$(x + \Delta x)^{n} \approx x^{n} + n(\Delta x)x^{n-1} \quad (\Delta x << 1) \quad (94)$$

Using this in Equation (93) gives

$$\frac{d(x^{n})}{dx} = \lim_{\Delta x \to 0} \left[ \frac{[x^{n} + n(\Delta x)x^{n-1}] - x^{n}}{\Delta x} \right]$$
(95)

We used an equal sign rather than an approximately equal sign in Equation (95) because our approximation formula (94) becomes exact when  $\Delta x$  becomes infinitesimally small.

In Equation (95), the terms  $x^n$  cancel and we are left with

$$\frac{d(x^{n})}{dx} = \lim_{\Delta x \to 0} \left[ \frac{n(\Delta x)x^{n-1}}{\Delta x} \right]$$
(96)

At this point, the factors  $\Delta x$  cancel and we have

$$\frac{d(x^{n})}{dx} = \lim_{\Delta x \to 0} \left[ nx^{n-1} \right]$$
(97)

Since no  $\Delta x$ 's remain in our formula, we end up with the exact result

$$\frac{d(x^n)}{dx} = nx^{n-1} \tag{98}$$

Equation (98) is the general formula for the derivative of the function  $x^n$ . In our discussion of integration, we saw that a constant could come outside the integral. The same thing happens with a derivative. Consider, for example,

$$\frac{d}{dx} \left[ a f(x) \right] = \lim_{\Delta x \to 0} \left| \frac{a f(x + \Delta x) - a f(x)}{\Delta x} \right|$$

Since the constant a has nothing to do with the limiting process, this can be written

$$\frac{d}{dx} \left[ af(x) \right] = a \lim_{\Delta x \to 0} \left[ \frac{f(x + \Delta x) - f(x)}{\Delta x} \right]$$

$$= a \frac{df(x)}{dx}$$
(99)

#### Exercise 5

Calculate the derivative with respect to x (i.e., d/dx) of the following functions. (When negative powers of x are involved, assume x is not equal to zero.)

- (a) x
- (b) x<sup>2</sup>
- (c) x<sup>3</sup>
- (d)  $5x^2 3x$

(Before you do part (d), use the definition of the derivative to prove that  $\frac{d}{dx}[f(x) + g(x)] = \frac{df(x)}{dx} + \frac{dg(x)}{dx}$ )

- (e) x<sup>-1</sup>
- (f) x<sup>-2</sup>
- (g) √x
- (h) 1/√x
- (i) 3x<sup>.73</sup>
- (j) 7x<sup>-.2</sup>
- (k) 1

(In part (k) first show that this should be zero from the definition of the derivative. Then write  $1 = x^0$  and show that Equation (98) also works, as long as x is not zero.)

(l) 5

10/ 1

### The Chain Rule

There is a simple trick called the *chain rule* that makes it easy to differentiate a wide variety of functions. The rule is

$$\frac{df[y(x)]}{dx} = \frac{df(y)}{dy}\frac{dy}{dx}$$
 chain rule (100)

To see how this rule works, consider the function

$$f(x) = (x^2)^n$$
(101)

We know that this is just  $f(x) = x^{2n}$ , and the derivative is

$$\frac{df(x)}{dx} = \frac{d}{dx}(x^{2n}) = 2nx^{2n-1}$$
(102)

But suppose that we did not know this trick, and therefore did not know how to differentiate  $(x^2)^n$ . We do, however, know how to differentiate powers like  $x^2$  and  $y^n$ . The chain rule allows us to use this knowledge in order to figure out how to differentiate the more complex function $(x^2)^n$ .

We begin by defining y(x) as

$$\mathbf{y}(\mathbf{x}) = \mathbf{x}^2 \tag{103}$$

Then our function  $f(x) = (x^2)^n can be written in terms of y as follows$ 

$$f(x) = (x^{2})^{n} = [y(x)]^{n} = (y)^{n} = f(y)$$
  
$$f(y) = (y)^{n}$$
 (104)

Differentiating (103) and (104) gives

$$\frac{\mathrm{d}y(x)}{\mathrm{d}x} = \frac{\mathrm{d}}{\mathrm{d}x}(x^2) = 2x \tag{105}$$

$$\frac{\mathrm{d}f(y)}{\mathrm{d}y} = \frac{\mathrm{d}}{\mathrm{d}y}(y^n) = ny^{n-1} \tag{106}$$

Using (104) and (105) in the chain rule (100) gives

$$\frac{df(y)}{dx} = \frac{df}{dy} \times \frac{dy}{dx} = (ny^{n-1}) \times (2x)$$
  
=  $2ny^{n-1}x$   
=  $2n(x^2)^{n-1}x$   
=  $2n(x^{2[n-1]})x$   
=  $2n(x^{[2n-2]})x$  (107)  
=  $2n(x^{[2n-2]+1})$   
=  $2nx^{2n-1}$ 

which is the answer we expect.

In our example, using the chain rule was more difficult than differentiating directly because we already knew how to differentiate  $x^{2n}$ . But we will shortly encounter examples of new functions that we do not know how to differentiate directly, but which can be written in the form f[y(x)], where we know df/dy and dy/dx. We can then use the chain rule to evaluate the derivative df/dx. We will give you practice with the chain rule when we encounter these functions.

#### **Remembering the Chain Rule**

The chain rule can be remembered by thinking of the dy's as cancelling as shown.

$$\frac{\mathrm{d}\mathbf{f}(\mathbf{y})}{\mathrm{d}\mathbf{y}}\frac{\mathrm{d}\mathbf{x}}{\mathrm{d}\mathbf{x}} = \frac{\mathrm{d}\mathbf{f}(\mathbf{y})}{\mathrm{d}\mathbf{x}} \qquad remembering \\ the chain rule \qquad (108)$$

## Partial Proof of the Chain Rule (optional)

The proof of the chain rule is closely related to cancellation we showed in Equation (108). A partial proof of the rule proceeds as follows.

Suppose we have some function f(y) where y is a function of the variable x. As a result f[y(x)] is itself a function of x and can be differentiated with respect to x.

$$\frac{d}{dx}f[y(x)] = \lim_{\Delta x \to 0} \frac{f[y(x + \Delta x)] - f[y(x)]}{\Delta x} \quad (123)$$

Now define the quantity  $\Delta y$  by

$$\Delta y \equiv y(x + \Delta x) - y(x) \tag{124}$$

so that

$$y(x + \Delta x) = y(x) + \Delta y$$
$$f[y(x + \Delta x)] = f(y + \Delta y)$$

and Equation (123) becomes

$$\frac{d}{dx}f[y(x)] = \lim_{\Delta x \to 0} \frac{f(y + \Delta y) - f(y)}{\Delta x}$$
(125)

Now multiply (125) through by

$$1 = \frac{\Delta y}{\Delta y} = \frac{y(x + \Delta x) - y(x)}{\Delta y}$$
(126)

to get

$$\begin{aligned} \frac{d}{dx} f[y(x)] \\ &= \lim_{\Delta x \to 0} \left[ \frac{f(y + \Delta y) - f(y)}{\Delta x} \times \frac{y(x + \Delta x) - y(x)}{\Delta y} \right] \\ &= \lim_{\Delta x \to 0} \left[ \frac{f(y + \Delta y) - f(y)}{\Delta y} \times \frac{y(x + \Delta x) - y(x)}{\Delta x} \right] \end{aligned}$$

(127)

where we interchanged  $\Delta x$  and  $\Delta y$  in the denominator.

(We call this a partial proof for the following reason. For some functions y(x), the quantity  $\Delta y = y(x + \Delta x) - y(x)$  may be identically zero for a small range of  $\Delta x$ . In that case we would be dividing by zero (the  $1/\Delta y$ ) even before we took the limit as  $\Delta x$  goes to zero. A more complete proof handles the special cases separately. The resulting chain rule still works however, even for these special cases.)

Since  $\Delta y = y(x + \Delta x) - y(x)$  goes to zero as  $\Delta x$  goes to zero, we can write Equation (127) as

$$\frac{d}{dx} f[y(x)]$$

$$= \begin{bmatrix} \liminf_{\Delta y \to 0} \frac{f(y + \Delta y) - f(y)}{\Delta y} \end{bmatrix}$$

$$\times \begin{bmatrix} \liminf_{\Delta x \to 0} \frac{y(x + \Delta x) - y(x)}{\Delta x} \end{bmatrix}$$

$$= \begin{bmatrix} \frac{df(y)}{dy} \end{bmatrix} \begin{bmatrix} \frac{dy}{dx} \end{bmatrix}$$
(100) repeated

This rule works as long as the derivatives df/dy and dy/dx are meaningful, i.e., we stay away from kinks or discontinuities in f and y.

# INTEGRATION FORMULAS

Knowing the formula for the derivative of the function  $x^n$ , and knowing that integration undoes differentiation, we can now use Equation (98)

$$\frac{\mathrm{d}x^{n}}{\mathrm{d}x} = nx^{n-1} \tag{98} repeated$$

to find the integral of the function  $x^n$ . We will see that this trick works for all cases except the special case where n = -1, i.e., the special case where the integral is a natural logarithm.

To integrate  $x^n$ , let us go back to our calculation of the distance  $s_x$  or x(t) traveled by an object moving in the x direction at a velocity  $v_x$ . This was given by Equations (19) or (56) as

$$\mathbf{x}(t)\Big|_{t_{\mathbf{i}}}^{\mathrm{T}} = \int_{t_{\mathbf{i}}}^{\mathrm{T}} \mathbf{v}_{\mathbf{x}}(t) \, \mathrm{d}t \tag{128}$$

where the instantaneous velocity  $v_x(t)$  is defined as

$$v_{x}(t) = \frac{dx(t)}{dt}$$
(129)

Suppose x(t) had the special form

$$\mathbf{x}(t) = t^{n+1} \qquad (a \ special \ case) \qquad (130)$$

then we know from our derivative formulas that

$$v(t) = \frac{dx(t)}{dt} = \frac{dt^{(n+1)}}{dt} = (n+1)t^n$$
 (131)

Substituting  $x(t) = t^{n+1}$  and  $v(t) = (n+1)t^n$  into Equation (128) gives

$$\mathbf{x}(t)\Big|_{t_{i}}^{\mathrm{T}} = \int_{t_{i}}^{\mathrm{T}} \mathbf{v}_{\mathbf{x}}(t) \, \mathrm{d}t \tag{128}$$

$$t^{n+1}\Big|_{t_{i}}^{T} = \int_{t_{i}}^{T} (n+1)t^{n}dt$$

$$= (n+1)\int_{t_{i}}^{T} t^{n}dt$$
(132)

Dividing through by (n+1) gives

$$\int_{t_{i}}^{T} t^{n} dt = \frac{1}{n+1} t^{n+1} \Big|_{t_{i}}^{T}$$
(133)

If we choose  $t_i = 0$ , we get the simpler result

$$\int_{0}^{T} t^{n} dt = \frac{T^{n+1}}{n+1}$$
(134)

and the indefinite integral can be written

$$\int t^{n} dt = \frac{t^{n+1}}{n+1}$$
 (135) (also 34)

This is the general rule we stated without proof back in Equation (34). Note that this formula says nothing about the case n = -1, i.e., when we integrate  $t^{-1} = 1/t$ , because n+1 = -1+1 = 0 and we end up with division by zero. But for all other values of n, we now have derived a general formula for finding the area under any curve of the form  $x^n$  (or  $t^n$ ). This is a rather powerful result considering the problems one encounters actually finding areas under curves. (If you did not do Exercise 1, the integration exercises on page 14, or had difficulty with them, go back and do them now.)

## **Derivative of the Exponential Function**

The previous work shows us that if we have a series expansion for a function, it is easy to obtain a formula for the derivative of the function. We will now apply this technique to calculate the derivative and integral of the exponential function  $e^x$ .

There is a series expansion for the function  $e^x$  that works for any value of x is but is most useful for small values of  $x = \alpha \ll 1$ , is

$$e^{\alpha} \approx 1 + \alpha + \frac{\alpha^2}{2!} + \frac{\alpha^3}{3!} + \cdots$$
 (136)

where  $2! = 2 \times 1$ ,  $3! = 3 \times 2 \times 1 = 6$ , etc. (The quantities 2!, 3! are called *factorials*. For example 3! is called *three factorial*.)

To see how well the series (136) works, consider the case  $\alpha = .01$ . From the series we have, up to the  $\alpha^3$  term

$$\begin{aligned} \alpha &= .01 \\ \alpha^2 &= .0001 ; & \alpha^2/2 &= .00005 \\ \alpha^3 &= .000001 ; & \alpha^3/6 &= .000000167 \end{aligned}$$

Giving us the approximate value

$$1 + \alpha + \frac{\alpha^2}{2!} + \frac{\alpha^3}{3!} = 1.010050167$$
(137)

When we enter .01 into a scientific calculator and press the  $e^x$  button, we get exactly the same result. Thus the calculator is no more accurate than including the  $\alpha^3$  term in the series, for values of  $\alpha$  equal to .01 or less.

Let us now see how to use the series (136) for calculating the derivative of  $e^x$ . We have, from the definition of a derivative,

$$\frac{d}{dx}f(x) \equiv \lim_{\Delta x \to 0} \frac{f(x + \Delta x) - f(x)}{\Delta x}$$
 (56) repeat

If  $f(x) = e^x$ , we get

$$\frac{d(e^{x})}{dx} = \lim_{\Delta x \to 0} \left[ \frac{e^{x + \Delta x} - e^{x}}{\Delta x} \right]$$
(138)

To do this calculation, we have to evaluate the quantity  $e^{x + \Delta x}$ . First, we use the fact that for exponentials

$$e^{a+b} = e^a e^b$$

(Remember that  $10^{2+3} = 10^2 \times 10^3 = 10^5$ .) Thus

$$e^{x + \Delta x} = e^x e^{\Delta x} \tag{139}$$

Now use the approximation formula (136), setting  $\alpha = \Delta x$  and throwing out the  $\alpha^2$  and  $\alpha^3$  and higher terms because we are going to let  $\Delta x$  go to zero

$$e^{\Delta x} \approx 1 + \Delta x \tag{140}$$

Substituting (140) in (139) gives

$$e^{x+\Delta x} \approx e^{x}(1+\Delta x)$$
$$= e^{x} + e^{x}\Delta x \qquad (141)$$

Next use (141) in (138) to get

$$\frac{d(e^{x})}{dx} = \lim_{\Delta x \to 0} \left[ \frac{(e^{x} + e^{x} \Delta x) - e^{x}}{\Delta x} \right]$$
(142)

The e<sup>x</sup> terms cancel and we are left with

$$\frac{d(e^{x})}{dx} = \lim_{\Delta x \to 0} \left[ \frac{e^{x} \Delta x}{\Delta x} \right] = \lim_{\Delta x \to 0} e^{x}$$
(143)

Since the  $\Delta x's$  cancelled, we are left with the exact result

$$\frac{\mathrm{d}(\mathrm{e}^{\mathrm{x}})}{\mathrm{d}\mathrm{x}} = \mathrm{e}^{\mathrm{x}} \tag{144}$$

We see that the exponential function  $e^x$  has the special property that it is its own derivative.

We will often want to know the derivative, not just of the function  $e^x$  but of the slightly more general result  $e^{ax}$  where a is a constant. That is, we want to find

$$\frac{\mathrm{d}}{\mathrm{dx}}\mathrm{e}^{\mathrm{ax}}$$
 (a = constant) (145)

Solving this problem provides us with our first meaningful application of the chain rule

$$\frac{df(y)}{dx} = \frac{df(y)}{dy}\frac{dy}{dx}$$
 (100) repeated

If we set

$$y = ax (146)$$

then we have

$$\frac{\mathrm{d}e^{\mathrm{ax}}}{\mathrm{d}x} = \frac{\mathrm{d}e^{\mathrm{y}}}{\mathrm{dy}}\frac{\mathrm{dy}}{\mathrm{dx}} \tag{147}$$

Now

$$\frac{\mathrm{d}\mathrm{e}^{\mathrm{y}}}{\mathrm{d}\mathrm{y}} = \mathrm{e}^{\mathrm{y}} \tag{148}$$

$$\frac{dy}{dx} = \frac{d}{dx}(ax) = a\frac{dx}{dx} = a \times 1 = a \qquad (149)$$

Using (148) and (149) in (147) gives

$$\frac{\mathrm{d}e^{\mathrm{ax}}}{\mathrm{dx}} = (\mathrm{e}^{\mathrm{y}})(\mathrm{a}) = (\mathrm{e}^{\mathrm{ax}})(\mathrm{a}) = \mathrm{a}\mathrm{e}^{\mathrm{ax}}$$

Thus we have

$$\frac{\mathrm{d}}{\mathrm{d}x}\mathrm{e}^{\mathrm{a}x} = \mathrm{a}\mathrm{e}^{\mathrm{a}x} \tag{150}$$

This result will be used so often it is worth memorizing.

# Exercise 6

For further practice with the chain rule, show that

$$\frac{de^{ax^2}}{dx} = 2axe^{ax^2}$$

Do this by choosing  $y = ax^2$ , and then do it again by choosing  $y = x^2$ .

# Integral of the Exponential Function

To calculate the integral of  $e^{ax}$ , we will use the same trick as we used for the integral of  $x^n$ , but we will be a bit more formal this time. Let us start with Equation (128) relating position x(t) and velocity v(t) = dx(t)/dt go get

$$x(t)\Big|_{t_{i}}^{t_{f}} = \int_{t_{i}}^{t_{f}} v_{x}(t) dt = \int_{t_{i}}^{t_{f}} \frac{dx(t)}{dt} dt \quad (128)$$

Since Equation (128) holds for any function x(t) [we did not put any restrictions on x(t)], we can write Equation (128) in a more abstract way relating any function f(x) to its derivative df(x)/dx

$$f(x)\Big|_{x_i}^{x_f} = \int_{x_i}^{x_f} \frac{df(x)}{dx} dx$$
(151)

To calculate the integral of  $e^{ax}$ , we set  $f(x) = e^{ax}$ and  $df(x)/dx = ae^{ax}$  to get

$$\left. e^{ax} \right|_{x_i}^{x_f} = \int_{x_i}^{x_f} a e^{ax} dx \tag{152}$$

Dividing (157) through by (a) gives us the definite integral

$$\int_{x_{i}}^{x_{f}} e^{ax} dx = \frac{1}{a} e^{ax} \Big|_{x_{i}}^{x_{f}} \qquad (a = constant) \quad (153)$$

The corresponding indefinite integral is

$$\int e^{ax} dx = \frac{e^{ax}}{a} \qquad (a = constant) \qquad (154)$$

## Exercise 7

The natural logarithm is defined by the equation

$$\ln(x) = \int \left(\frac{1}{x}\right) dx$$
 (see Equations 33-40)

Use Equation (151) to show that

$$\frac{\mathrm{d}}{\mathrm{d}x}(\ln x) = \frac{1}{x} \tag{155}$$

(Hint—integrate both sides of Equation (155) with respect to x.)

#### **DERIVATIVE AS THE SLOPE OF A CURVE**

Up to now, we have emphasized the idea that the derivative of a function f(x) is given by the limiting process

$$\frac{df(x)}{dx} = \lim_{\Delta x \to 0} \left[ \frac{f(x + \Delta x) - f(x)}{\Delta x} \right]$$
 (55) repeat

We saw that this form was convenient when we had an explicit way of calculating  $f(x + \Delta x)$ , as we did by using a series expansion. However, a lot of words are required to explain the steps involved in doing the limiting process indicated in Equation (55). In contrast, the idea of an integral as being the area under a curve is much easier to state and visualize. Now we will provide an easy way to state and interpret the derivative of a curve.

Consider the function f(x) graphed in Figure (20). At a distance x down the x axis, the curve had a height f(x) as shown. Slightly farther down the x axis, at  $x + \Delta x$ , the curve has risen to a height  $f(x + \Delta x)$ .



Two points on a curve, a distance  $\Delta x$  apart.



**Figure 20a** At this point, the curve is tilted by approximately an angle  $\theta^*$ .

Figure (20a) is a blowup of the curve in the region between x and  $x + \Delta x$ . If the distance  $\Delta x$  is sufficiently small, the curve between x and  $x + \Delta x$ should be approximately a straight line and that part of the curve should be approximately the hypotenuse of the right triangle abc seen in Figure (20a). Since the side opposite to the angle  $\theta^*$  is  $f(x + \Delta x) - f(x)$ , and the adjacent side is  $\Delta x$ , we have the result that the tangent of the angle  $\theta^*$  is

$$\tan(\theta^*) = \frac{f(x + \Delta x) - f(x)}{\Delta x}$$
(156)

When we make  $\Delta x$  smaller and smaller, take the limit as  $\Delta x \rightarrow 0$ , we see that the angle  $\theta^*$  becomes more nearly equal to the angle  $\theta$  shown in Figure (21), the angle of the curve when it passes through the point x. Thus

$$\tan \theta = \lim_{\Delta x \to 0} \frac{f(x + \Delta x) - f(x)}{\Delta x}$$
(157)

The tangent of the angle at which the curve passes through the point x is called the *slope of the curve at the point x*. Thus from Equation (157) we see that the slope of the curve is equal to the derivative of the curve at that point. We now have the interpretation that the derivative of a curve at some point is equal to the slope of the curve at that point, while the integral of a curve is equal to the area under the curve up to that point.





The tangent of the angle  $\theta$  at which the curve passes through the point x is called the slope of the curve at that point.

# **Negative Slope**

In Figure (22) we compare the slopes of a rising and a falling curve. In (22a), where the curve is rising, the quantity  $f(x + \Delta x)$  is greater than f(x) and the derivative or slope

$$\frac{\mathrm{d}f(x)}{\mathrm{d}x} = \lim_{\Delta x \to 0} \left| \frac{f(x + \Delta x) - f(x)}{\Delta x} \right|$$

is a positive number.

In contrast, for the downward curve of Figure (22b),  $f(x + \Delta x)$  is less than f(x) and the slope is negative. For a curve headed downward, we have

$$\frac{df(x)}{dx} = -\tan(\theta) \qquad \frac{downward heading}{curve} \quad (158)$$

(For this case you can think of  $\theta$  as a negative angle, so that  $\tan(\theta)$  would automatically come out negative. However it is easier simply to remember that the slope of an upward directed curve is positive and that of a downward directed curve is negative.)



**Figure 22 a,b** Going uphill is a positive slope, downhill is a negative slope.

#### Exercise 8

Estimate the numerical value of the slope of the curve shown in Figure (23) at points (a), (b), (c), (d) and (e). In each case do a sketch of  $[f(x + \Delta x) - f(x)]$  for a small  $\Delta x$ , and let the slope be the ratio of  $[f(x + \Delta x) - f(x)]$  to  $\Delta x$ . Your answers should be roughly 1, 0, -1, + $\infty$ ,  $-\infty$ .



**Figure 23** *Estimate the slope at the various points indicated.* 

# THE EXPONENTIAL DECAY

A curve that we will encounter several times during the course is the function  $e^{-ax}$  shown in Figure (24), which we call an exponential decay. Since exponents always have to be dimensionless numbers, we are writing the constant (a) in the form  $1/x_0$  so that the exponent  $x/x_0$  is more obviously dimensionless.

The function  $e^{-x/x_0}$  has several very special properties. At x = 0, it has the numerical value 1 ( $e^0 = 1$ ). When we get up to  $x = x_0$ , the curve has dropped to a value

$$e^{-x/x_0} = e^{-1} = \frac{1}{e} \quad (at \ x = x_0)$$
  
 $\approx \frac{1}{2.7}$ 
(159)

When we go out to  $x = 2x_0$ , the curve has dropped to

$$e^{-2x_0/x_0} = e^{-2} = \frac{1}{e^2}$$
 (160)

Out at  $x = 3x_0$ , the curve has dropped by another factor of e to (1/e)(1/e)(1/e). This decrease continues indefinitely. It is the characteristic feature of an exponential decay.

## Muon Lifetime

In the muon lifetime experiment, we saw that the number of muons surviving decreased with time. At the end of two microseconds, more than half of the original 648 muons were still present. By 6 micro-



#### Figure 24

As we go out an additional distance  $x_0$ , the exponential curve drops by another factor of 1/e.

seconds, only 27 remained. The decay of these muons is an example of an exponential decay of the form

number of  
surviving muons = 
$$\begin{pmatrix} number of \\ muons at \\ time t = 0 \end{pmatrix} \times e^{-t/t_0}$$
 (161)

where  $t_0$  is the time it takes for the number of muons remaining to drop by a factor of 1/e = 1/2.7. That time is called the muon *lifetime*.

We can use Equation (161) to estimate the muon lifetime  $t_0$ . In the movie, the number of muons at the top of the graph, reproduced in Figure (25), is 648. That is at time t = 0. Down at time t = 6 microseconds, the number surviving is 27. Putting these numbers into Equation (161) gives

$$27 \frac{\text{surviving}}{\text{muons}} = 648 \frac{\text{initial}}{\text{muons}} \times e^{-6/t_0}$$
$$e^{-6/t_0} = \frac{27}{648} = .042$$
(162)

Take the natural logarithm *ln* of both sides of Equation (162), [remembering that  $ln(e^x) = x$ ] gives

$$ln(e^{-6/t_0}) = \frac{-6}{t_0} = ln(.042) = -3.17$$

where we entered .042 on a scientific calculator and pressed the ln key. Solving for  $t_0$  we get

$$t_0 = \frac{6 \text{ microsec}}{3.17} = 1.9 \text{ microseconds} \quad (163)$$

This is close to the accepted value of  $t_0 = 2.20$  microseconds which has been determined from the study of many thousands of muon decays.



#### Figure 25

The lifetime of each detected muon is represented by the length of a vertical line. We can see that many muons live as long as 2 microseconds  $(2\mu s)$ , but few live as long as 6 microseconds.

# **Half Life**

The exponential decay curve  $e^{-t/t_0}$  decays to 1/e = 1/2.7 of its value at time  $t_0$ . While 1/e is a very convenient number from a mathematical point of view, it is easier to think of the time  $t_{1/2}$  it takes for half of the muons to decay. This time  $t_{1/2}$  is called the *half life* of the particle.

From Figure (26) we can see that the half life  $t_{1/2}$  is slightly shorter than the lifetime  $t_0$ . To calculate the half life from  $t_0$ , we have

$$e^{-t/t_0}\Big|_{t=t_{1/2}} = e^{-t_{1/2}/t_0} = \frac{1}{2}$$
 (164)

Again taking the natural logarithm of both sides of Equation (164) gives

$$ln(e^{-t_{1/2}/t_0}) = \frac{-t_{1/2}}{t_0} = ln(\frac{1}{2}) = -.693$$

$$\boxed{t_{1/2} = .693 t_0}$$
(165)

From Equation (165) you can see that a half life  $t_{1/2}$  is about .7 of the lifetime  $t_0$ . If the muon lifetime is 2.2µsec (we will abbreviate microseconds as µsec), and you start with a large number of muons, you would expect about half to decay in a time of

$$(t_{1/2})_{muon} = .693 \times 2.2 \mu sec = 1.5 \mu sec$$

The basic feature of the exponential decay curve  $e^{-t/t_0}$  is that for every time  $t_0$  that passes, the curve decreases by another factor of 1/e. The same applies to the half life  $t_{1/2}$ . After one half life,  $e^{-t/t_0}$  has decreased to half its value. After a second half life, the curve is down to  $1/4 = 1/2 \times 1/2$ . After 3 half lives it is down to  $1/8 = 1/2 \times 1/2 \times 1/2$  as shown in Figure (27).



To help illustrate the nature of exponential decays, suppose that you started with a million muons. How long would you expect to wait before there was, on the average, only one left?

To solve this problem, you would want the number  $e^{-t/t_0}$  to be down by a factor of 1 million

$$e^{-t/t_0} = 1 \times 10^{-6}$$

Taking the natural logarithm of both sides gives

$$ln(e^{-t/t_0}) = \frac{-t}{t_0} = ln(1 \times 10^{-6}) = -13.8 \quad (166)$$

(To calculate  $ln(1 \times 10^{-6})$ , enter 1, then press the *exp* key and enter 6, then press the +/- key to change it to -6. Finally press = to get the answer -13.8.)

Solving Equation (166) for t gives

t = 
$$13.8 t_0 = 13.8 \times 2.2 \,\mu \text{sec}$$
  
t = 30 microseconds (167)

That is the nature of an exponential decay. While you have nearly half a million left after around 2 microseconds, they are essentially all gone by 30 microseconds.

#### **Exercise 9**

How many factors of 1/2 do you have to multiply together to get approximately 1/1,000,000? Multiply this number by the muon half-life to see if you get about 30 microseconds.



After each half-life, the curve decreases by another factor of 1/2.

**Figure 26** (1/2)Comparison of the lifetime  $t_0$  and the half-life  $t_{1/2}$ .

# Measuring the Time Constant from a Graph

The idea that the derivative of a curve is the slope of the curve, leads to an easy way to estimate a lifetime  $t_0$  from an exponential decay curve  $e^{-t/t_0}$ .

The formula for the derivative of an exponential curve is

$$\frac{\mathrm{d}\mathrm{e}^{\mathrm{a}\mathrm{t}}}{\mathrm{d}\mathrm{t}} = \mathrm{a}\mathrm{e}^{\mathrm{a}\mathrm{t}} \tag{150} \text{ repeated}$$

Setting  $a = -1/t_0$  gives

$$\frac{d}{dt}\left(e^{-t/t_0}\right) = -\frac{1}{t_0}e^{-t/t_0}$$
(168)

Since the derivative of a curve is the slope of the curve, we set the derivative equal to the tangent of the angle the curve makes with the horizontal axis.

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\mathrm{e}^{-t/t_0}\right) = -\frac{1}{t_0}\mathrm{e}^{-t/t_0} = \tan\theta \qquad (168a)$$

The minus sign tells us that the curve is headed down.

In Figure (28), we have drawn a line tangent to the curve at the point t = T. This line intersects the (t) axis (the axis where  $e^{-t/t_0}$  goes to zero) at a distance (x) down the *t* axis.



#### Figure 28

A line, drawn tangent to the exponential decay curve at some point T, intersects the axis a distance x down the axis. We show that this distance x is equal to the time constant  $t_0$ . This is true no matter what point T we start with.

The height (y) of the point where we drew the tangent curve is just the value of the function  $e^{-T/t_0}$ . The tangent of the angle  $\theta$  is the opposite side (y) divided by the adjacent side (x)

$$\tan\theta = \frac{y}{x} = \frac{e^{-T/t_0}}{x}$$
(169)

Equating the two magnitudes of  $\tan \theta$  in Equations (169) in (168a) gives us

$$\frac{1}{t_0}e^{-T/t_0} = \frac{1}{x}e^{-T/t_0}$$

which requires that

$$\mathbf{x} = \mathbf{t}_0 \tag{170}$$

Equation (170) tells us that the distance (x), the distance down the axis where the tangent lines intersect the axis, is simply the time constant  $t_0$ .

The result gives us a very quick way of determining the time constant  $t_0$  of an exponential decay curve. As illustrated in Figure (29), choose any point on the curve, draw a tangent to the curve at that point and measure the distance down the axis where the tangent line intersects the axis. That distance will be the time constant  $t_0$ . We will use this technique in several laboratory exercises later in the course.



A quick way to estimate the time constant  $t_0$  for an exponential decay curve is to draw the tangent line as shown.

# THE SINE AND COSINE FUNCTIONS

The final topic in our introduction to calculus will be the functions  $\sin\theta$  and  $\cos\theta$  and their derivatives and integrals. We will need these functions when we come to rotational motion and wave motion.

The definition of  $\sin\theta$  and  $\cos\theta$ , which should be familiar from trigonometry, are

$$\sin\theta = \frac{a}{c} \left(\frac{opposite}{hypotenuse}\right)$$
 (171a)

$$\cos\theta = \frac{b}{c} \left(\frac{adjacent}{hypotenuse}\right)$$
 (171b)



where  $\theta$  is an angle of a right triangle as shown in Figure (30), (a) is the length of the side opposite to  $\theta$ , (b) the side adjacent to  $\theta$  and (c) the hypotenuse.

The formulas are simplified if we consider a right triangle whose hypotenuse is of length c = 1 as in Figure (31). Then we have

$$\sin\theta = a$$
 (172a)

 $\cos\theta = b \tag{172b}$ 



We can then fit our right triangle inside a circle of radius 1 as shown in Figure (32).



**Figure 32** *Fitting our right triangle inside a unit radius circle.* 

## **Radian Measure**

We are brought up to measure angles in degrees, but physicists and mathematicians usually measure angles in *radians*. The angle  $\theta$  measured in radians is defined as the *arc length*  $\ell$  subtended by the angle  $\theta$  on a circle of unit radius, as shown in Figure (32).

$$\theta_{\text{radians}} = \ell \qquad \begin{array}{c} arc \ length \ subtended \\ by \ \theta \ on \ a \ unit \ circle \end{array}$$
(173)

(If we had a circle of radius c, then we would define  $\theta_{radians} = \ell/c$ , a dimensionless ratio. In the special case c = 1, this reduces to  $\theta_{radians} = \ell$ .)

Since the circumference of a unit circle is  $2\pi$ , we see that  $\theta$  for a complete circle is  $2\pi$  radians, which is the same as 360 degrees. This tells us how to convert from degrees to radians. We have the conversion factor

$$\frac{360 \text{ degrees}}{2\pi \text{ radians}} = 57.3 \frac{\text{degrees}}{\text{radian}}$$
(174)

As an example of using this conversion factor, suppose we want to convert 30 degrees to radians. We would have

$$\frac{30 \text{ degrees}}{57.3 \text{ degrees/radian}} = .52 \text{ radians} \quad (175)$$

To decide whether to divide by or multiply by a conversion factor, use the dimensions of the conversion factor. For example, if we had multiplied 30 degrees by our conversion factor, we would have gotten

$$30 \text{ degrees} \times 57.3 \frac{\text{degrees}}{\text{radian}} = 1719 \frac{\text{degrees}^2}{\text{radian}}$$

This answer may be correct, but it is useless.

The numbers to remember in using radians are the following:

$$90^\circ = \pi/2$$
 radians  
 $180^\circ = \pi$  radians  
 $270^\circ = 3\pi/2$  radians  
 $360^\circ = 2\pi$  radians
(176)

The other values you can work out as you need them.

# **The Sine Function**

In Figure (33) we have started with a circle of radius 1 and, in a somewhat random way, labeled 10 points around the circle. The arc length up to each of these points is equal to the angle, in radian measure, subtended by that point. The special values are:

$$\theta_0 = 0 \text{ radians}$$
  

$$\theta_4 = \pi/2 \text{ radians } (90^\circ)$$
  

$$\theta_6 = \pi \text{ radians } (180^\circ)$$
  

$$\theta_8 = 3\pi/2 \text{ radians } (270^\circ)$$
  

$$\theta_{10} = 2\pi \text{ radians } (360^\circ)$$

In each case the  $\sin\theta$  is equal to the height (a) at that point. For example

$$sin\theta_1 = a_1$$
  

$$sin\theta_2 = a_2$$
  

$$....$$
  

$$sin\theta_{10} = a_{10}$$

We see that the height (a) starts out at  $a_0 = 0$  for  $\theta_0$ , increases up to  $a_4 = 1$  at the top of the circle, drops back down to  $a_6 = 0$  at  $\theta_6 = \pi$ , goes negative, down to  $a_8 = -1$  at  $\theta_8 = 3\pi/2$ , and returns to  $a_{10} = 0$  at  $\theta_{10} = 2\pi$ . Our next step is to construct a graph in which  $\theta$  is shown along the horizontal axis, and we plot the value of  $\sin\theta = (a)$  on the vertical axis. The result is shown in Figure (34). The eleven points, representing the heights  $a_0$  to  $a_{10}$  at  $\theta_0$  to  $\theta_{10}$  are shown as large dots in Figure (34). We have also sketched in a smooth curve through these points, it is the curve we would get if we had plotted the value of (a) for every value of  $\theta$  from  $\theta = 0$  to  $\theta = 2\pi$ . The smooth curve is a graph of the function *sin*  $\theta$ .

# Exercise 10

Using the fact that the cosine function is defined as

 $\cos \theta = b$  (b is defined in Figures 31, 32)

plot the values of  $b_0, b_1, \dots, b_{10}$  on a graph similar to Figure (34), and show that the cosine function  $\cos \theta$  looks like the curve shown in Figure (35).



Figure 33 The heights a; at various points around a unit circle.

**Figure 34** Graph of the function  $sin(\theta)$ .

There is nothing that says we have to stop measuring the angle  $\theta$  after we have gone around once. On the second trip around,  $\theta$  increases from  $2\pi$  up to  $4\pi$ , and the curve  $\sin\theta$  repeats itself. If we go around several times, we get a result like that shown in Figure (36). We often call that a *sine wave*.

Several cycles of the curve  $\cos\theta$  are shown in Figure (37). You can see that the only difference between a sine and a cosine curve is where you set  $\theta = 0$ . If you move the origin of the cosine axis back (to the left) 90° ( $\pi/2$ ), you get a sine wave.

Figure 35 The cosine function.



csinθ

A graph of the function  $y(\theta) = c \sin \theta$  looks just like the curve in Figure (36), except the curve goes up to a height c and down to -c as shown in Figure (38). We would get the curve of Figure (38) by plotting points around a circle as in Figure (33), but using a circle of radius c. We call this factor c the amplitude of the sine wave. The function  $\sin\theta$  has an amplitude 1, while the sine wave in Figure (38) has an *amplitude c* (its values range from +c to -c).



Figure 37 Several cycles of the curve  $cos(\theta)$ .

# **Derivative of the Sine Function**

Since the sine and cosine functions are smooth curves, we should be able to calculate the derivatives and integrals of them. We will do this by first calculating the derivative, and then turning the process around to find the integral, just as we did for the functions  $x^n$  and  $e^x$ .

The derivative of the function  $\sin\theta$  is defined as usual by

$$\frac{\mathrm{d}(\sin\theta)}{\mathrm{d}\theta} = \lim_{\Delta\theta\to0} \left| \frac{\sin(\theta + \Delta\theta) - \sin\theta}{\Delta\theta} \right| \quad (177)$$

where  $\Delta \theta$  is a small change in the angle  $\theta$ .

The easiest way to evaluate this limit is to go back to the unit circle of Figure (25) and construct both  $\sin\theta$ and  $\sin(\theta + \Delta\theta)$  as shown in Figure (39). We see that  $\sin\theta$  is the height of the triangle with an angle  $\theta$ , while  $\sin(\theta + \Delta\theta)$  is the height of the triangle whose center angle is  $(\theta + \Delta\theta)$ . What we have to do is calculate the difference in heights of these two triangles.

In Figure (40) we start by focusing our attention on the slender triangle abc with an angle  $\Delta \theta$  at (a) and long sides of length 1 (since we have a unit circle). Since the angle  $\Delta \theta$  is small, the short side of this triangle is essentially equal to the arc length along the circle from point (b) to point (c). And since we are using radian measure, this arc length is equal to the angle  $\Delta \theta$ .



Triangles for the sin $\theta$  and the sin ( $\theta + \Delta \theta$ ).

Now draw a line vertically down from point (c) and horizontally over from point (b) to form the triangle bcd shown in Figure (40). The important point is that the angle at point (c) in this tiny triangle is the same as the angle  $\theta$  at point (a). To prove this, consider the sketch in Figure (41). A line bf is drawn tangent to the circle at point (b), so that the angle abf is a right angle. That means the other two angles in the triangle add up to 90°, the total angle in any triangle being 180°

$$\theta + \varphi = 90^{\circ} \tag{178}$$

Since the angle at (e) in triangle bef is also a right angle, the other two angles in the triangle bef, must also add up to  $90^{\circ}$ .

$$\alpha + \varphi = 90^{\circ} \tag{179}$$

For both Equations (178) and (179) to be true, we must have  $\alpha = \theta$ .



The difference between  $\sin\theta$  and  $\sin(\theta + \Delta\theta)$  is equal to the height of the side cd of the triangle cdb.



Demonstration that the angle  $\alpha$  equals the angle  $\theta$ .

The final step is to note that when  $\Delta \theta$  in Figure (40) is very small, the side cb of the very small triangle is essentially tangent to the circle, and thus parallel to the side bf in Figure (41). As a result the angle between cb and the vertical is also the same angle  $\theta$ .

Because the tiny triangle, shown again in Figure (42) has a hypotenuse  $\Delta \theta$  and a top angle  $\theta$ , the vertical side, which is equal to the difference between  $\sin \theta$  and  $\sin(\theta + \Delta \theta)$  has a height  $(\cos \theta)\Delta \theta$ . Thus we have

$$\sin(\theta + \Delta \theta) - \sin \theta = (\cos \theta) \Delta \theta \qquad (180)$$

н

Equation (180) becomes exact when  $\Delta \theta$  becomes an infinitesimal angle.

We can now evaluate the derivative

$$\frac{\mathrm{d}(\sin\theta)}{\mathrm{d}\theta} = \lim_{\Delta\theta\to0} \left[ \frac{\sin(\theta + \Delta\theta) - \sin\theta}{\Delta\theta} \right]$$
$$= \lim_{\Delta\theta\to0} \left[ \frac{(\cos\theta)\Delta\theta}{\Delta\theta} \right]$$
$$= \lim_{\Delta\theta\to0} \cos\theta$$

Thus we get the exact result

$$\frac{d}{d\theta}(\sin\theta) = \cos\theta \qquad (181)$$

$$\frac{d}{d\theta}(\sin\theta) = \cos\theta \qquad (181)$$

$$\frac{d}{\theta}(\sin\theta) = \cos\theta \qquad (181)$$

$$\frac{d}{\theta}(\sin\theta) = \cos\theta \qquad (181)$$

#### Figure 42

The difference between  $\sin\theta$  and  $\sin(\theta + \Delta\theta)$  is equal to  $\Delta\theta\cos\theta$ . Check that this result is reasonable by considering the special cases  $\theta = 0$  and  $\theta = 90^{\circ} (\pi/2)$ .

#### Exercise 11

Using a similar derivation, show that

$$\frac{\mathrm{d}}{\mathrm{d}\theta}(\cos\theta) = -\sin\theta \tag{182}$$

#### Exercise 12

Using the chain rule for differentiation, show that

$$\begin{vmatrix} \frac{d}{d\theta} (\sin a\theta) &= a \cos a\theta \\ \frac{d}{d\theta} (\cos a\theta) &= -a \sin a\theta \end{vmatrix} (a = constant)$$
(183)

(Hint—if you need to, look at Equation (145) through (150).

#### Exercise 13

Using the fact that integration reverses differentiation, as we did in integrating the function  $e^x$  (Equations (151) through (154), show that

$$\int_{\theta_{i}}^{\theta_{f}} (\cos a\theta) d\theta = \frac{1}{a} \sin a\theta \Big|_{\theta_{i}}^{\theta_{f}}$$
(184a)  
$$\int_{\theta_{i}}^{\theta_{f}} (\sin a\theta) d\theta = -\frac{1}{a} \cos a\theta \Big|_{\theta_{i}}^{\theta_{f}}$$
(184b)

Use sketches of the integrals from  $\theta_i = 0$  to  $\theta_f = \pi/2$  to show that Equations (184a) and (184b) have the correct numerical sign. (Explicitly explain the minus sign in (184b).

# Calculus 2000-Chapter 2 Second Derivatives and the One Dimensional Wave Equation

In our discussion of a wave pulse on a rope, in Chapter 15 of Physics 2000, we used a combination of physical observation and a somewhat tricky argument to show that the speed of the wave pulse was given by the formula  $v = \sqrt{T/\rho}$ . The physical observation was noting that a pulse travels down the rope at an apparently uniform speed. The trick was to analyze the behavior of the rope from the point of view of someone moving along with the pulse (as on pages 15-4, 5).

Another way to handle the problem is to directly apply Newton's second law to a section of the rope. When we use this direct approach, we end up with an equation that involves second derivatives not only with respect in time, but also with respect to space. The resulting equation with its second derivatives is what is known as the **wave equation**. The aim of this chapter is to learn how to handle the wave equation, at least for waves moving in one dimension. (Handling three dimensional wave equations comes later.)

To use the wave equation with any real understanding, not just manipulating formulas, requires more familiarity with the properties of a second derivative than we have needed so far. Thus we will begin this chapter with a discussion of the second derivative, and then go on to the one dimensional wave equation.

# THE SECOND DERIVATIVE

We have already encountered the idea of a *second derivative* in our discussion of velocity and acceleration. Consider a particle moving down the x axis, whose position is described by the function x(t). The particle's velocity  $v_x(t)$  is given by

$$v_x(t) = \frac{dx(t)}{dt}$$
 first (1)

which is the first derivative, with respect to time, of x(t). The particle's acceleration a(t) is given by

$$a_{x}(t) = \frac{dv_{x}(t)}{dt}$$
(2)

When we use (1) for v(t) in Equation (2) we get

$$a_{x}(t) = \frac{d}{dt} \left[ \frac{dx(t)}{dt} \right] \qquad \substack{second \\ derivative} \qquad (3)$$

In Equation (3), we see that  $a_x(t)$  is obtained from x(t) by differentiating twice with respect to time. We say that  $a_x(t)$  is the second derivative of x(t) and use the simplified notation

$$\frac{d}{dt} \left[ \frac{dx(t)}{dt} \right] \equiv \frac{d^2 x(t)}{dt^2} \qquad \begin{array}{c} simplified \\ notation for \\ second derivative \end{array} (4)$$

With this notation, the position x(t), velocity  $v_x(t)$ , and acceleration  $a_x(t)$  are related by

$$v_{x}(t) = \frac{dx(t)}{dt}$$

$$a_{x}(t) = \frac{d^{2}x(t)}{dt^{2}}$$
(5)

There is nothing particularly difficult about carrying out a second derivative, just do the derivative operation twice as illustrated in the following example.

# Example 1

Calculate the second derivative, with respect to  $\theta$  , of  $sin(a\theta)$ 

$$\frac{d^2 \sin(a\theta)}{d\theta^2} = ? \tag{6}$$

#### Solution

Begin by taking the first derivative

$$\frac{d\sin(a\theta)}{d\theta} = a\cos(a\theta) \tag{7}$$

Now differentiate again

$$\frac{d}{d\theta} \left( \frac{d \sin(a\theta)}{d\theta} \right) = \frac{d}{d\theta} \left[ a \cos(a\theta) \right]$$
$$= a \frac{d}{d\theta} \left[ \cos(a\theta) \right]$$
$$= a \left[ -a \sin(a\theta) \right]$$
(8)

Thus we get

$$\frac{d^2 \sin(a\theta)}{d\theta^2} = -a^2 \sin(a\theta)$$
(9)

We see that the second derivative of a sine curve is itself a sine curve, with a minus sign.

# Exercise 1

Calculate the following second derivatives

(a) 
$$\frac{d^{2}}{d\theta^{2}} \left[ \cos(a\theta) \right]$$
  
(b) 
$$\frac{d^{2}}{dx^{2}} \left[ e^{-ax} \right]$$
  
(c) 
$$\frac{d^{2}}{dx^{2}} \left[ \ln(x) \right]$$
  
(d) 
$$\frac{d^{2}}{dx^{2}} \left[ \ln(ax) \right]$$
  
(e) 
$$\frac{d^{2}}{dx^{2}} \left[ x^{n} \right]$$
  
(f) 
$$\frac{d^{2}}{dx^{2}} \left[ (ax)^{n} \right]$$
  
(g) 
$$\frac{d^{2}}{dx^{2}} \left[ \frac{1}{(x)^{n}} \right]$$
  
(h) 
$$\frac{d^{2}}{dx^{2}} \left[ \frac{1}{(ax)^{n}} \right]$$
# Geometrical Interpretation of the Second Derivative

We have seen that the various calculus operations have a geometrical interpretation. Integration was equivalent to finding the area under a curve, while the first derivative represented the slope of a curve. We now want to obtain the geometrical interpretation of the second derivative. We will see that the second derivative is equal to what we will call the *curvature* of the curve. To see exactly what that is, consider the following derivation.

Let y(x) be the section of a circle as shown in Figure (1). Let us use notation found in a number of calculus texts, and denote the derivative of y(x) by y'(x)

$$y'(x) \equiv \frac{dy(x)}{dx}$$
 simplified  
notation (10)

In terms of y'(x), the second derivative is

$$\frac{d^2 y(x)}{dx^2} = \lim_{\Delta x \to 0} \frac{y'(x + \Delta x) - y'(x)}{\Delta x}$$
(11)

Remember that y'(x) = dy/dx is the slope of the curve at position x as shown in Figure (2) (For example, see Figure 21 of Chapter 1). Thus Equation (11) tells us that to find the second derivative of y(x) we have to find the change in slope as we move from x to  $x + \Delta x$ .



Figure 1

Calculating the change in the slope of the circle, as we go from x = 0 to  $x = \Delta x$ .



We will evaluate the second derivative at the bottom of the circle, where the curve is horizontal and the slope is zero.

$$\mathbf{y'}[\mathbf{x}=\mathbf{0}] = 0 \qquad \begin{array}{c} curve \ horizontal\\ at \ x=0 \end{array} \tag{12}$$

Now move down the x axis a distance  $\Delta x$  as shown in Figure (1). If  $\Delta x$  is small, then  $\Delta x$  is essentially equal to the arc length  $\Delta \ell$  along the circle, and the angle  $\Delta \theta$  in radian measure is the arc length divided by the radius R of the circle

$$\Delta \theta = \frac{\Delta \ell}{R} \approx \frac{\Delta x}{R} \tag{13}$$

If we draw a line tangent to the circle at position  $x = \Delta x$ , this tangent line will make an angle  $\Delta \theta$  to the horizontal as shown in Figure (1). (The two angles labeled  $\Delta \theta$  in Figure 1 are equal no matter how big  $\Delta \theta$  is.) Thus the slope of the tangent line at  $x = \Delta x$  is

slope of circle = 
$$tan(\Delta\theta)$$
 (14)

Now if  $\Delta \theta$  is a small angle, which it will be if we take the limit as  $\Delta x \rightarrow 0$ , we can use the approximation

$$\tan\left(\Delta\theta\right) \approx \Delta\theta \tag{15}$$

You can see why this approximation is good for small angles from Figure (2a).

Thus the slope of the tangent line at  $x = \Delta x$  is given by

slope of  
tangent line = 
$$y'[x = \Delta x] = \Delta \theta = \frac{\Delta x}{R}$$
 (16)  
at  $x = \Delta x$ 

where we used Equation (13) for  $\Delta \theta$ .

Now we have values of y' at x = 0 (Equation 12) and at  $x = \Delta x$  (Equation 16), we can use these values in Equation (10) to get the value of  $d^2y/dx^2$  at x = 0, i.e., at the bottom of the circle.



Introducing the notation

$$\frac{d^2 y(x)}{dx^2}\Big|_{x=0} \qquad \begin{array}{c} means \ d^2 y(x)/dx^2 \\ evaluated \ at \ the \\ point \ x=0 \end{array}$$

We have from Equation (10)

$$\frac{d^2 y(x)}{dx^2} \bigg|_{x=0} = \lim_{\Delta x \to 0} \frac{y'[x = \Delta x] - y'[x = 0]}{\Delta x}$$
(17)

With  $y'[x=\Delta x] = \Delta x/R$  (Equation 16) and y'[x=0] = 0, we get

$$\frac{d^2 y(x)}{dx^2} \bigg|_{x=0} = \lim_{\Delta x \to 0} \frac{\Delta x/R - 0}{\Delta x}$$
$$= \lim_{\Delta x \to 0} \frac{1}{R}$$

Since the  $\Delta x$ 's canceled, we see that 1/R is the limiting value and we get

$$\left. \frac{\mathrm{d}^2 \mathbf{y}(\mathbf{x})}{\mathrm{d}\mathbf{x}^2} \right|_{\mathbf{x}=0} = \frac{1}{\mathrm{R}} \tag{18}$$

With a slightly messier derivation we could calculate  $d^2y/dx^2$  anywhere around the circle, not just at the bottom, and we get the same answer 1/R. Thus we have the more general result



# CURVATURE

Consider the curve shown in Figure (3) representing some function y(x). At point  $x_0$  we have drawn a circle that just fits against the curve. The radius of the circle is adjusted to give the closest match possible between the curve  $y(x_0)$  and the circle. When we get this closest fit, both the first and the second derivatives of the circle and y(x) are equal at  $x = x_0$ . In other words

$$\left. \frac{\mathrm{d}^2 \mathbf{y}(\mathbf{x})}{\mathrm{d}\mathbf{x}^2} \right|_{\mathbf{x} = \mathbf{x}_0} = \frac{1}{\mathrm{R}}$$
(20)

In Figure (3) the quantity R is called the *radius of curvature* of the curve y(x) at the point  $x_0$ , and 1/R is called the *curvature* 

$$\frac{1}{R} \equiv curvature \text{ of the curve}$$
(21)

You can see intuitively why 1/R is called curvature. If R is very large, the curve is almost flat and we would say it has little curvature. As R becomes smaller, the curve bends in a tighter circle, and the curvature 1/R becomes greater.

This is the geometrical picture of the second derivative. While the first derivative was equal to the slope of the curve at some point, the *second derivative is equal to the curvature* of the curve at that point. The curvature is explicitly the reciprocal of the radius of curvature of the curve where the radius of curvature is found by fitting a circle to the curve as in Figure (3). [*Exercise: under what circumstances would the second derivative or curvature be negative?*]



# **Figure 3** At any point along a curve, the curvature is 1/R or -1/R, where R is the radius of the circle that just fits the curve as shown.

# **Curve Fitting and Boat Lofting**

The problem of working with curves has a number of practical applications, one of the more interesting of which, at least to a sailor, is the *lofting* of boats.

It turns out that the eye is extremely good at judging the smoothness of a curve. We can, for example, easily spot the slightest wrinkle in what is supposed to be the smooth side of a boat. (It is an interesting question as to how the eye and brain can do this so well.)

Through the 16th century, boats were rather crude looking. Starting in the 17th century, better looking boats were built using the following steps. The first was to carve a model of the hull that was to be built. Then conceptually slice the model as you would slice a loaf of bread. Each of these cuts was called a *station*. Typically one used about 15 stations, each representing a cross section of the hull at different distances along the length of the boat. Then points were taken from the model to represent the shape of the hull at each station. Figure (4) is a typical example of a hull cross section at a middle station.

Since the points showing the shape of the hull were taken from a small model, any errors in measurement would be greatly magnified when the hull was laid out full scale. An error of a fraction of a millimeter in measuring the model would lead to a very obvious bump in the final hull shape.



Typical cross section. (Since boats are supposed to be symmetric, only one side is usually drawn.)

To avoid these bumps, the plans were taken up into the loft of the boat shed (hence the name *lofting*), and drawn full scale. Wooden splines, typically thin strips of spruce, were bent along the points of the curve. Since the splines bent along smooth curves, any points that were out of place would not be fitted by the spline and the points would be moved to fit the smooth curve. This process is called *spline fitting*. Once all the full scale curves were smoothed by spline fitting, then the boat hull was constructed using these smoothed plans and the result, if done correctly, was a smooth, good looking hull.

In the early 1970's, shortly after we had started using the computer in teaching introductory physics, we had lunch with a boat builder who described the rather tedious process of lofting a boat. He wondered if lofting could be done more easily on the computer. This was before the availability of inexpensive line plotters, so that the work would all have to be done numerically. We agreed to try, the incentive being a reduced price on a diesel engine for our boat if we successfully lofted the boat builder's new lobster boat design.

The most successful part of the project was finding an easy and very effective way to spot a smooth curve. Just print out a list of the third derivatives of the curve. Since the second derivative is the curvature of the curve, the third derivative is the rate at which the curvature is changing as you go along the curve. If the curvature changes slowly, then the curve looks smooth. A bump represents a sudden change in curvature and therefore has a large third derivative. What a spruce spline essentially does is to minimize the third derivative.

About the same time that we wrote the lofting program, a physicist, Peter Karos in Germany, also wrote a boat lofting program. As one does not make much of a living from a lofting program, Karos turned to the problem of using the computer to create letter forms. The letters of the alphabet are constructed from different curves that depend upon which font you are using. And just as in boat design, the eye is very sensitive to the smoothness of the curves, even for relatively small letters. Karos based his boat lofting and letter design programs on what are called **Besier** curves. To construct a Besier curve through a series of points, at each point you specify the location y(x) of the point, the first derivative y'(x) = dy/dx, and the second derivative  $y''(x) = d^2y/dx^2 = 1/R$ . The section of curve between two adjacent points is then constructed to match the first and second derivative at the end points and minimize the third derivative in the region in between. This uniquely determines the line.

Karos's techniques using Besier curves was built into the Postscript<sup>TM</sup> language used for letter design. A way of graphically handling the construction of Besier curves was developed and became the basis of the *Adobe Illustrator*<sup>TM</sup> program.

Those of you who have used Adobe Illustrator, or any of the similar drawing programs, will be familiar with the constructing of Besier curves. You place the pen tool at a point and press the mouse button. That establishes the point y(x). Then you drag the pen tool in some direction. That direction establishes the slope of the curve y'(x) at that point. How far out you drag the pen tool before you let up on the mouse button determines the radius of curvature R at that point, and thus establishes the second derivative  $y''(x) = d^2y/dx^2 = 1/R$  there (see Figure 5). When you move the mouse to another point, press the mouse button and drag, you determine y(x), y'(x)and y''(x) at the new point, and then the computer draws the smooth Besier curve between the two points.

When you are using Adobe Illustrator, or other drawing programs, think of the fact that you are controlling the position, the first derivative, and the second derivative every time you place and drag the mouse.

#### Figure 5

Constructing Besier curves with Adobe Illustrator®. In that program, the radius of curvature is set to about 60% of the distance that the cursor is pulled out from the point.



# THE BINOMIAL EXPANSION

We have seen, starting in Chapter 1 of the Physics text, the usefulness of the binomial expansion

$$(1 + \alpha)^n = 1 + n\alpha + \frac{n(n-1)}{2!}\alpha^2 + \cdots$$
 (22)

which is valid for any value of  $\alpha$  less than one, but which gets better as  $\alpha$  becomes smaller. For very small  $\alpha$ , we could neglect all terms involving  $\alpha^2$  or higher powers of  $\alpha$ , giving us the approximation formula

$$(1+\alpha)^n \approx 1+n\alpha \qquad (\alpha <<1)$$
 (23)

which is good for any value of n.

With calculus, we can easily derive the formula for the various terms in the binomial expansion. We begin with the assumption that the quantity  $(1 + \alpha)^n$ can be expanded in some kind of a series involving powers of  $\alpha$ . We will write the series in the form

$$(1 + \alpha)^n = A_0 \alpha^0 + A_1 \alpha^1 + A_2 \alpha^2 + A_3 \alpha^3 + \cdots$$
(24)

where the  $A_0$ ,  $A_1$ ,  $A_2$ , etc. are unknown coefficients that we have to determine.

Equation (24) is supposed to be correct for small values of  $\alpha$  including  $\alpha = 0$ . Setting  $\alpha = 0$  gives

$$(1+0)^{n} = A_{0}0^{0} + A_{1}0^{1} + A_{2}0^{2} + A_{3}0^{3} + \cdots$$
(25)

Here is a peculiar convention we use. We assume that any number  $x^0 = 1$  no matter what x is, including  $0^0$ . Thus  $A_0 0^0 = A_0$ , all the other terms on the right side of Equation (25) are zero, and we get

$$1^{n} = 1 = A_{0}$$
 (26)

which determines  $A_0$ .

(Writing  $A_0 \alpha^0$  instead of just  $A_0$  for the first term in the series is formalism that makes the series look more consistent, but is unnecessary if you do not like the idea of  $0^0 = 1$ .) To determine the value of  $A_1$ , differentiate Equation (24) with respect to  $\alpha$ . We get, using the chain rule

$$\frac{\mathrm{d}}{\mathrm{d}\alpha}(1+\alpha)^{n} = \left[\frac{\mathrm{d}}{\mathrm{d}(1+\alpha)}(1+\alpha)^{n}\right] \left[\frac{\mathrm{d}(1+\alpha)}{\mathrm{d}\alpha}\right]$$
$$= \left[n(1+\alpha)^{n-1}\right] \left[1\right]$$
$$= n(1+\alpha)^{n-1} \qquad (27)$$

Differentiating the right hand side of Equation (24) gives

$$\frac{\mathrm{d}}{\mathrm{d}\alpha} \left( A_0 \alpha^0 + A_1 \alpha^1 + A_2 \alpha^2 + A_3 \alpha^3 + \cdots \right)$$
$$= 0 + A_1 + 2A_2 \alpha + 3A_3 \alpha^2 + \cdots$$
(28)

Thus the first derivative of Equation (24), with respect to  $\alpha$ , is

$$n(1 + \alpha)^{n-1} = A_1 + 2A_2\alpha + 3A_3\alpha^2 + \cdots$$
 (29)

Now set  $\alpha = 0$  and we get

$$n(1+0)^{n-1} = A_1 + 2A_2 \times 0 + 3A_3 \times 0^2 + \cdots$$
 (30)

which gives us

$$\mathbf{n} = \mathbf{A}_1 \tag{31}$$

and determines the coefficient  $A_1$ .

To determine  $A_2$ , differentiate Equation (29) with respect to  $\alpha$ . With

$$\frac{d}{d\alpha}(1+\alpha)^{n-1} = (n-1)(1+\alpha)^{n-2}$$

we get

$$n\left[(n-1)(1+\alpha)^{n-2}\right] = 2A_2 + 3(2\alpha) + \cdots (32)$$

Setting  $\alpha = 0$  gives

$$A_2 = \frac{n(n-1)}{2}$$
(33)

#### Exercise 2

Differentiate Equation (32) with respect to  $\alpha$ , set  $\alpha = 0$ , and show that A<sub>3</sub> is given by

$$A_3 = \frac{n(n-1)(n-2)}{3 \times 2 \times 1}$$
(34)

From Equation (34) you can see the general formula emerging

$$A_{n} = \frac{n(n-1)(n-2)(n-3)\cdots}{n!}$$
(35)

Thus by successive differentiation we can rather easily determine all the terms in the binomial expansion.

(One thing we have not worried about, but which is of major concern in calculus texts, is the range of values of  $\alpha$  for which the series is valid. Such questions are important from a purely mathematical point of view, but are seldom of practical importance. From a practical point of view, you can usually evaluate a few terms, and if the last ones are negligibly small, the series is probably good enough.)

# The Taylor Series Expansion

The binomial expansion we have just discussed is a special case of the more general expansion called the *Taylor series expansion*. In Figure (6) we have sketched a curve representing some function

$$\mathbf{y} = \mathbf{f}(\mathbf{x}) \tag{36}$$

Suppose we know everything about the function at the point  $x_0$  and would like to figure out where the curve is going as we move away from that point. By knowing everything about f(x) at the point  $x_0$ , we mean that we know  $f(x_0)$  as well as all the derivatives of f(x) evaluated at  $x = x_0$ .



#### Figure 6

If we know everything about the curve y = f(x)at the point  $x_0$ , can we predict where the curve will be a short distance farther down the x axis? The derivation of the Taylor series expansion begins with the assumption that the function f(x) can be expanded, in the vicinity of the point  $x_0$  by the so called *power series* 

$$f(x) = A_0(x - x_0)^0 + A_1(x - x_0)^1 + A_2(x - x_0)^2 + \cdots$$
(37)

If you think of  $(x - x_0)$  as being some small distance  $\alpha$ , then the expansion in Equation (37) is the same form as the expansion of the function  $(1 + \alpha)^n$  back in Equation (24). The difference is that for different functions f(x) we get different coefficients  $A_n$ .

To calculate the  $A_n$ , we do the same thing that we did in deriving the binomial expansion. We differentiate both sides of the equation and then set  $x = x_0$  (which corresponds to setting  $\alpha = (x - x_0)$  equal to zero).

First we set  $x = x_0$  in Equation (37) to get

$$f(x_0) = A_0(x - x_0)^0 + A_1(x - x_0)^1 + \cdots$$
  
=  $A_0(0)^0 + A_1(0)^1 + \cdots$  (38)  
=  $A_0$ 

which determines the first coefficient  $A_0$ .

Differentiating both sides of Equation (37) with respect to x and then setting  $x = x_0$  gives

$$f'(x) \equiv \frac{df(x)}{dx} = A_1 + 2A_2(x - x_0) + 3A_3(x - x_0)^3 + \cdots$$
(39)

where we used the chain rule to show that

$$\frac{d}{dx}(x - x_0)^n = n(x - x_0)^{n-1}$$
(40)

Setting  $x = x_0$  in Equation (39) gives

$$f'(x_0) \equiv \left. \frac{df(x)}{dx} \right|_{x = x_0} = A_1 \tag{41}$$

all the other terms being zero.

#### Exercise 3

Show that

$$A_{2} = \frac{1}{2} \left. \frac{d^{2}f(x)}{dx^{2}} \right|_{x = x_{0}} = \frac{f''(x_{0})}{2}$$
(42)

$$A_{3} = \frac{1}{3 \times 2 \times 1} \left. \frac{d^{3}f(x)}{dx^{3}} \right|_{x = x_{0}} = \frac{f'''(x_{0})}{3!}$$
(43)

From Exercise 3 you can see that the general form of the Taylor series expansion is

$$f(x - x_0) = f(x_0) + f'(x_0)(x - x_0)^1 + \frac{1}{2!}f''(x_0)(x - x_0)^2 + \frac{1}{3!}f'''(x_0)(x - x_0)^3 + \cdots$$

This can be written in the compact form

$$f(x - x_0) = \sum_{n=0}^{\infty} \frac{f^n(x_0)}{n!} (x - x_0)^n \qquad Taylor series expansion$$
(44)

where we used the notation

$$f^{n}(x_{0}) \equiv \left. \frac{d^{n}f(x)}{dx^{n}} \right|_{x = x_{0}}$$

$$\tag{45}$$

The tricky part of the mathematics of the Taylor series expansion is how far you can go, how far x can be away from  $x_0$ , and still have a valid expansion. Perhaps more important to the physicist is how far you can go before you have to include too many terms and the expansion is not useful.

# Exercise 4

Apply the Taylor series expansion, Equation (44) to the function

$$f(x) = (x - x_0)^{n}$$

evaluated at  $x_0 = 1$ , and show that you get the binomial expansion. (Hint—set  $\alpha = x - x_0$ , i.e., substitute  $x = x_0 + \alpha$  at the end.)

# **The Constant Acceleration Formulas**

While the Taylor series expansion, Equation (44), looks like a very new topic, we have been using a Taylor series expansion since the very beginning of our discussion of calculus. The constant acceleration formulas are a simple example of this expansion.

Figure (7) is a reproduction of our instantaneous velocity drawing, Figure (3-32d) from Chapter 3 of the Physics text and Figure (Cal 1-1d) of the Calculus text. At some instant of time, the ball is located at some position  $(x_0, y_0)$  at time  $t_0$ , and we wish to predict the position of the ball at some later time t.

The location of the ball is described by two functions x(t) and y(t). We know  $x(t_0)$ ,  $y(t_0)$  and all the derivatives of these functions at time  $t_0$ , they are simply the velocity and acceleration

$$\mathbf{x}'(t) = \frac{\mathrm{d}\mathbf{x}(t)}{\mathrm{d}t} = \mathbf{v}_{\mathbf{x}}(t) \tag{46}$$

$$x''(t) = \frac{d^2 x(t)}{dt^2} = a_x(t)$$
 (47)

$$y'(t) = \frac{dy(t)}{dt} = v_y(t)$$
(48)

$$y''(t) = \frac{d^2 y(t)}{dt^2} = a_y(t)$$
 (49)



**Figure 7** *Instantaneous velocity at time (t).* 

If the particle is moving with constant acceleration, then all higher derivatives are zero. For example

$$y'''(t) \equiv \frac{d^{3}y(t)}{dt^{3}} = \frac{da_{y}(t)}{dt} = 0 \quad for \\ constant \\ acceleration$$
(50)

The Taylor series expansion for y directed motion y(t) is

$$y(t - t_0) = \sum_{n=0}^{\infty} y^n(t_0) \frac{(t - t_0)^n}{n!}$$
  
=  $y_0(t_0) + \frac{dy(t)}{dt} \Big|_{t_0} \frac{(t - t_0)}{1!}$   
 $+ \frac{d^2 y(t)}{dt^2} \Big|_{t_0} \frac{(t - t_0)^2}{2!} + \cdots$ 

With  $dy/dt = v_y$  and  $d^2y/dt^2 = a_y$ , we get

$$y(t - t_0) = y_0(t_0) + v_y(t_0)(t - t_0) + \frac{1}{2}a_y(t_0)(t - t_0)^2$$
(51)

with all higher powers of  $(t - t_0)$  having zero coefficients.

If we set  $t_0 = 0$  Equation (51), we get the very familiar result

$$y(t) = y_0 + v_{y0}t + \frac{1}{2}a_yt^2$$
 (52)

Here is an example of a Taylor series expansion that is valid for any range of values  $(t - t_0)$ . It is good for all times t because all derivatives of y(t) above the second derivative are zero.)

### Exercise 5

Suppose a particle is moving in the y direction with a constantly increasing acceleration. I.e., assume that

$$a'_{y}(t) \equiv \frac{da_{y}(t)}{dt} = constant$$

Find the formula for y(t) for all future times t. (This is one step above the constant acceleration formulas.)

# THE WAVE EQUATION

In the Physics text, we calculated the speed of a wave pulse on a rope in Chapter 15, pages 15-4 and 5. As we mentioned in the introduction, the calculation was relatively simple because of two tricks we were able to pull. One was to walk along with the pulse, so that it looked as if the pulse were standing still and the rope were passing through it. The second was to picture the top of the pulse as an arc of a circle, so that we would know the acceleration of the rope as it went around the arc. We got the right answer, but the process did not generate much confidence that we could handle more general cases, like calculating the speed of a sound wave pulse, or even of a compressional pulse on a Slinky. (Remember that we used dimensional analysis, an important but approximate tool, to estimate the wave speeds in these cases.)

What we will do now is the more direct approach of applying Newton's laws to a section of the wave pulse, get a differential equation, which happens to involve second derivatives in both space and time, and then solve the differential equation in the usual way. That is, we guess a solution, plug it into the equation, and see if we made the correct guess. We will use as much physical insight as we can to guide us in making the guess. The differential equation we will be working with is called the *wave equation*.

Here we will be working with the wave equation for waves moving in one dimension. The three dimensional wave equation will be discussed later.



Wave pulse on a rope. The curvature is positive (points up) at  $x_1$ , and negative at  $x_2$ .

# Waves on a Rope

Our analysis of a wave pulse on a rope begins much as it did in Chapter 15. Figures (8) and (9) are similar to Figures (15-3c) and (15-3d), except that we are now standing still relative to the rope, and we are assuming the pulse is passing by us.

In our current analysis of the wave pulse, we will be somewhat more formal than we were in Chapter 15. We will say that the rope, at the present time, lies along a curve y(x) as shown in Figure (8). The quantity x is the distance down the rope (say from one end) and y(x) is the height of the pulse there, i.e., the distance the rope is displaced from its equilibrium position. From our various discussions of derivatives, we know that dy(x)/dx is the slope of the rope at position x, and  $d^2y(x)/dx^2 = 1/R(x)$  is the curvature, which is equal to the reciprocal of the radius of curvature R(x) at that point. In Figure (8) we have sketched in circles to show the radius of curvature at the two points  $x_1$  and  $x_2$  along the curve. The curvature is positive at  $x_1$  and negative at  $x_2$ .

Let us consider a short section of rope of length  $\Delta l$ located at position x as shown in Figure (9). For now assume that this section begins at the top of the pulse where the rope is horizontal. Shortly we will see that our results apply at any position along the rope.

The two ends of the section of rope are being pulled along the rope by the tension T. If the rope were straight, if there were no curvature at this point, the tension forces would cancel each other and there would be no net force on  $\Delta \ell$ . Only because there is curvature is there a net force which we have labeled  $T_v$  in Figure (9).



Due to the tension pulling on both sides, this section of rope feels a net downward force  $T_y \approx T\Delta\theta$ .

As long as  $\Delta \ell$  is short enough, this section of rope will lie along the circle we have drawn to show the radius of curvature, and the two tension forces  $\vec{T}$ will be tangent to the circle at the two ends. The result, from geometry we have now seen several times, is that the two angles labeled  $\Delta \theta$  will be equal and the right hand tension force will have a downward pointing component  $T_y$  given by

$$T_v = T \sin(\Delta \theta) \approx T \Delta \theta$$
 (53)

where for small angles we can replace the sine of an angle by the angle itself.

From Figure (9) we see that the angle  $\Delta \theta$  is given by

$$\Delta \theta = \frac{\Delta \ell}{R} \tag{54}$$

so that

$$T_{y} = T\Delta\theta = T\Delta\ell\left(\frac{1}{R}\right)$$
(55)

Since 1/R is the curvature  $d^2y(x)/dx^2$  at  $\Delta \ell$ , we get

$$T_{y}(x) = T\Delta \ell \frac{d^{2}y(x)}{dx^{2}}$$
(56)

While Equation (56) was derived starting from the top of the pulse, we can see that as long as the sides of the pulse are not steep, as long as we are dealing with a shallow wave pulse, Equation (56) should apply all along the wave.

To see this, we have in Figure (10) analyzed the net force  $\vec{T}_y$  acting slightly to the left side of the top of the pulse (at point  $x_2$  in Figure (8)). Actually Figure (10) is the same as Figure (9), rotated by an angle  $\varphi = dy(x)/dx$  which is the slope of the rope at point  $x_2$ . Here is where the shallow wave approximation comes in. As long as the wave is shallow and the



Figure 10

If the section of rope slopes at an angle  $\varphi$ , then the net force  $T_y'$  slopes at the same angle. That has little effect as long as the waves are shallow and  $\varphi$  remains small.

sides of the pulse do not become steep, the angle  $\varphi$  will be small, there will be very little rotation of Figure (9), and the net force  $\vec{T}_{y'}$  will point nearly straight down and have a magnitude close to that given by Equation (56).

On the other hand, if the pulse becomes steep, the net force is no longer y directed and our current analysis will no longer apply. Whoever has watched ocean waves break as they approach the beach and become steeper and steeper, will recognize that steep waves behave very differently from shallow ones. Here we are working only with the theory of shallow waves.

Returning to Equation (56), which we have written here again

$$T_y(x) = T\Delta \ell \frac{d^2 y(x)}{dx^2}$$
 (56) repeated

we want to point out that this equation gives us not only the magnitude but also the direction of the net force  $T_y$ . Where the curvature  $d^2y(x)/dx^2$  is positive, as it is at point  $x_1$  in Figure (8), the net force  $T_y$ is directed upwards. Where the curvature is negative as at point  $x_2$ , the net force  $T_y$  points down. Thus Equation (56) for  $T_y(x)$  correctly changes sign when the direction of the net force changes.

Now that we have a reasonably general formula for the net force  $T_y$  on a section  $\Delta \ell$  of the rope (the only approximation being the shallow wave approximation), we are ready to apply Newton's second law, relating this net force to the mass  $\Delta m$  and the acceleration  $\vec{a}(t)$  of this section.

If the rope has a mass density  $\mu$  kg/meter, then the mass of a section of length  $\Delta \ell$  is simply

$$\Delta m = \mu \Delta \ell \qquad \frac{mass of}{section \,\Delta \ell} \tag{57}$$

We need to think a bit more about the situation to describe the acceleration of  $\Delta m$ . So far we have described the rope by the curve y(x), which is essentially a single snapshot of the rope at some special time t.

Another way to look at a wave pulse is to look at one point on the rope, and watch the point move up and down as the pulse comes by. We can describe this changing height by the function y(t). The acceleration  $a_v(t)$  is then given by

$$a_{y}(t) = \frac{d^{2}y(t)}{dt^{2}}$$
(58)

Equation (58) is limited in that it describes the motion of only one point of the rope. We can describe the motion of the whole rope for all times with a function y(x,t) that is a function of both space and time. If we look at the rope at some instant of time  $t_0$ , then the shape of the rope is given by

$$y(x) = y(x,t) \Big|_{t=t_0}$$
 (59)

while if we stand at one point  $\mathbf{x}_0$ , the motion of the rope is given by

$$y(t) = y(x,t) \Big|_{x = x_0}$$
 (60)

An explicit example of such a function y(x,t) was our traveling wave formula of Equation (15-26) of the Physics text

$$y(x,t) = A \sin(kx - \omega t)$$
 (61) (also 15-26)

which as we saw represented a sinusoidal wave traveling to the right at a speed

$$v_{wave} = \frac{\omega}{k}$$
 (62) (also 15-30)

where the special frequency k is related to the wavelength  $\lambda$  by  $k = 2\pi/\lambda$ , and the angular frequency  $\omega$  is related to period T by  $\omega = 2\pi/T$ . (As a quick exercise, show that  $\omega/k$  has the dimensions of a velocity).

With Equation (61) for y(x,t), you can easily see that if you look at the wave at one time, say t = 0, then

$$y(x,t)\Big|_{t=0} = y(x) = \sin(kx)$$
 (63)

is a pure spacial sine wave. If you look at one particular point, for instance, x = 0, you get

$$y(x,t)\Big|_{x=0} = y(t) = \sin(-\omega t)$$
 (64)

which is a pure sinusoidal oscillation.

# **Partial Derivatives**

When dealing with a function of two or more variables, like y(x,t), we have to be somewhat careful when we talk about derivatives. For now, we will always assume that if we are differentiating with respect to space, we will hold the time variable constant, i.e., consider the curve at one instant of time. Conversely, if we are differentiating with respect to time, we will consider only one point in space, i.e., hold x constant. There is a special notation for these so called *partial derivatives*, where we differentiate with respect to one variable holding the other constant. In this notation we replace the d's, as in dx or dt by the symbol  $\partial$ . Thus

$$\frac{\mathrm{d}y(\mathbf{x},t)}{\mathrm{d}\mathbf{x}}\Big|_{\mathrm{holding t constant}} \equiv \frac{\mathrm{d}y(\mathbf{x},t)}{\mathrm{d}\mathbf{x}} \tag{65}$$

$$\frac{\mathrm{d}y(x,t)}{\mathrm{d}x}\Big|_{\mathrm{holding } x \mathrm{ constant}} \equiv \frac{\partial y(x,t)}{\partial t}$$
(66)

With this notation we get, for

$$y(x,t) = \sin(kx - \omega t)$$
 (67a)

$$\frac{\partial y(x,t)}{\partial x} = k \cos (kx - \omega t)$$
 (67b)

$$\frac{\partial y(\mathbf{x},t)}{\partial t} = -\omega \cos\left(\mathbf{k}\mathbf{x} - \omega t\right) \tag{67c}$$

Using this new notation for partial derivatives, our Equations (56) for the net force  $T_y$  on  $\Delta \ell$ , and (58) for the acceleration  $a_y$  of  $\Delta \ell$  becomes

$$T_{y}(x,t) = T\Delta \ell \frac{\partial^{2} y(x,t)}{\partial x^{2}}$$
(56a)

$$a_{y}(x,t) = \frac{\partial^{2} y(x,t)}{\partial t^{2}}$$
(58a)

To apply Newton's second law, we equate the net force  $T_y(x,t)$  to the mass  $\Delta m = \mu \Delta \ell$  times the acceleration  $a_v(x,t)$  to get

$$T_{y}(x,t) = \Delta m a_{y}(x,t)$$
$$T\Delta \ell \frac{\partial^{2} y(x,t)}{\partial x^{2}} = (\mu \Delta \ell) \frac{\partial^{2} y(x,t)}{\partial t^{2}}$$
(68)

The factors of  $\Delta \ell$  cancel, and after dividing through by  $\mu$  we get

$$\left| \frac{T}{\mu} \frac{\partial^2 y(x,t)}{\partial x^2} = \frac{\partial^2 y(x,t)}{\partial t^2} \right|$$
(69)

as our final differential equation for the motion of the wave pulse on the rope.

How do you solve such a differential equation? As we have mentioned several times, we guess an answer for y(x,t), and plug the guess into the differential equation to see if we have made the correct guess. Also, we use whatever physics we have available to help us make a good guess.

Right now we do not have a formula for a single pulse that we can use as a guess for a solution to Equation (69). However we do have the formula in Equation (61) for a sine wave traveling to the right at a speed  $v = \omega/k$ 

$$y(x,t) = A \sin(kx - \omega t)$$
 (61) repeated

To see if this traveling wave is a solution to our differential Equation (69), we have to take a number of partial derivatives. They are

$$\frac{\partial y(x,t)}{\partial x} = \frac{\partial}{\partial x} A \sin(kx - \omega t)$$
$$= A k \cos(kx - \omega t)$$
(70a)

$$\frac{\partial^2 y(x,t)}{\partial x^2} = \frac{\partial}{\partial x} Ak \cos(kx - \omega t)$$
$$= -A k^2 \sin(kx - \omega t)$$
(70b)

$$\frac{\partial y(x,t)}{\partial t} = \frac{\partial}{\partial t} A \sin(kx - \omega t)$$
$$= -\omega A \cos(kx - \omega t)$$
(70c)

$$\frac{\partial^2 y(x,t)}{\partial t^2} = \frac{\partial}{\partial t} (-\omega A) \cos(kx - \omega t)$$
$$= -(-\omega A)(-\omega) \sin(kx - \omega t)$$
(70d)

Using Equations (70b) and (70d) in Equation (69) gives

$$\frac{T}{\mu} (-Ak^2) \sin(kx - \omega t) \stackrel{?}{=} -A\omega^2 \sin(kx - \omega t)$$
(71)

The question mark in Equation (71) means that this is a guess, and we still have to see if the guess works.

First we notice that the functions  $\sin(kx - \omega t)$  cancel. We had to have this cancellation or there was no chance of making the two sides equal for all times t and all positions x. We also note that the amplitudes A cancel, which means that the solution does not depend upon the amplitude A. After these cancellations we get

$$\frac{T}{\mu}(-k^2) = -\omega^2$$

$$\frac{T}{\mu} = \frac{\omega^2}{k^2} = v_{wave}^2$$
(72)

where we noted that  $v_{wave} = \omega/k$ . Taking the square root of Equation (72) gives

$$v_{wave} = \sqrt{\frac{T}{\mu}}$$
(73)

which is the answer we got in the Physics text, Equation (15-5), for the speed of a pulse on a rope.

# **The One Dimensional Wave Equation**

If we go back to Equation (69), and replace T/ $\!\mu\,$  by  $v_{wave}^{}^{2}$  , we get

$$v_{wave}^2 \frac{\partial^2 y(x,t)}{\partial x^2} = \frac{\partial^2 y(x,t)}{\partial t^2}$$
 one dimensional wave equation

(74)

This is a general form of what is called the *one dimensional wave equation*. As we have just seen, a traveling sine wave, moving to the right at a speed  $v_{wave}$  is a solution to this equation. The following exercises demonstrate that waves traveling to the left, and standing waves, are also solutions to this equation.

# **Exercise 6**

(a) The formula for a sine wave moving to the left at a speed  $v_{wave} = \omega/k$  was given in Equation (15-33) of the Physics text as

 $y(x,t)_{wave moving left} = A \sin(kx + \omega t)$  (15-33)

Show that this wave also obeys the wave Equation (73).

(b) Later in Chapter 15 we saw that a standing wave, which is the sum of a left moving and a right moving traveling wave, was given by the formula

$$y = A \sin kx \cos \omega t$$
 (15-35)

Show that this wave is also a solution to the wave Equation.

# Exercise 7

Suppose you have two solutions  $y_1(x,t)$  and  $y_2(x,t)$ , both of which are a solution to the wave equation with the same speed  $v_{wave}$ . Show that the sum wave

$$y(x,t) = y_1(x,t) + y_2(x,t)$$
(75)

is also a solution of the same wave equation.

Exercise 7 gives us an important result. For our wave equation, which we got by considering wave pulses that were not too steep, the sum of two or more waves, each of which is a solution of the wave equation, is itself a solution.

In our discussion of Fourier analysis, introduced on page 16-6 of the Physics text, we saw that any continuous curve can be constructed from a sum of sine wave shapes. This suggests that we could construct a single wave pulse, moving to the left at a speed v<sub>wave</sub>, by adding up a bunch of traveling sine waves of different wavelengths  $\lambda_i = 2\pi/k_i$ , but all with the same speed v<sub>wave</sub> =  $\omega_i/k_i$ . The construction in Figure (11) suggests how we could add the sine (actually cosine) waves to get a pulse. Since each wave is a solution to the same wave Equation (73), the sum, i.e., the single pulse, is also a solution.

From Figure (11), it should be clear that we can construct a solution to the wave equation representing a pulse with very steep sides. However, in our analysis of the motion of the rope, we had to restrict ourselves to shallow waves in order to derive the wave equation for pulses on the rope. What this means is that the wave equation has solutions that we will not see on the rope. The shallow pulses on the rope will obey the wave equation, but we should expect that a steep pulse on the rope will behave differently. Not as differently as a breaking ocean wave, but differently.

#### Figure 11

How to add cosine waves to get a pulse. At x = 0, all the waves add to give a big amplitude y. As we go out from x = 0, there is more and more cancellation until the sum wave adds to zero. If all these are traveling waves moving to the right at the same speed  $v_{wave} = \omega_i k_i$ , then the whole pulse must move at the same speed, maintaining its shape.



# **Compressional Waves on a Spring**

When we came to the discussion of compressional waves on a spring, in particular the compressional Slinky wave we saw in Figure (Phys1-6) reproduced here, we resorted to dimensional analysis in Chapter 15 of the Physics text because there are no obvious tricks to calculate the speed of the pulse. Now we are in a position to set up a differential equation describing the motion of a short segment  $\Delta \ell$  of the spring. We will get the wave equation, and from that we can immediately tell the speed of the pulse.

Suppose we have a stretched spring of length L as shown in Figure (12). The force required to stretch the spring, which is equal to the tension T in the spring, is given by Hook's law as

$$\mathbf{T} = \mathbf{k}(\mathbf{L} - \mathbf{L}_0) \tag{76}$$

where  $L_0$  is the unstretched length of the spring.

Now suppose that we stretch the spring an additional amount  $\Delta L$ . The tension will increase by an amount  $\Delta T$  given by

$$T + \Delta T = k(L + \Delta L - L_0)$$
$$= k(L - L_0) + k\Delta L$$

Using Equation (76) to cancel the T and  $k(L - L_0)$  terms, we are left with

$$\Delta T = k\Delta L = kL\left(\frac{\Delta L}{L}\right)$$
(77)



Figure 1-6 (Physics 2000) Compressional wave on a Slinky.

Figure 12 A tension T stretches the spring from a length  $L_0$ to a length L.



There are two reasons why we have written  $\Delta T$  as  $kL(\Delta L/L)$  rather than just  $k\Delta L$ . The first is that  $\Delta L/L$  is the amount of *stretch per unit length*, a quantity engineers call *strain*. It is a more inherent property of the spring than the total stretch  $\Delta L$ .

The second reason is that the product kL is also an inherent property of the spring. In Chapter 15, page 15-7 of the Physics text, we saw that if you had two identical springs of spring constant k, and attached them together, you got a spring twice as long but with half the spring constant. It is the product kL that does not change when you connect identical springs or cut a spring in half. Engineers would call this inherent property kL of the spring a *spring modulus*.

To describe the stretched spring, we will introduce a function y(x) that represents the displacement of a point on the spring from its equilibrium (or initial) position. When we stretch a spring from a length L to a length  $L + \Delta L$ , as shown in Figure (13), every point on the spring moves to the right a distance y(x)given by the formula

$$\mathbf{y}(\mathbf{x}) \begin{pmatrix} displacement \\ of \ a \ point \\ on \ the \ spring \end{pmatrix} = \frac{\mathbf{x}}{\mathbf{L}} \Delta \mathbf{L}$$
(78)

where x is the distance down the spring, starting at the left end. You can see where we got Equation (78). If we are at the left end where x = 0, y(x) = 0and there is no displacement. At the right end, where x = L, we get the full displacement  $y(L) = (L/L)\Delta L = \Delta L$ . In Equation (78) we are assuming that the displacement increases uniformly as we go down the spring.



# Figure 13

The displacement y(x) increases as we go down the spring. With the formula  $y(x) = (x/L)\Delta L$ , we are assuming that the displacement is increasing uniformly. If we differentiate y(x) with respect to x we get

$$\frac{dy(x)}{dx} = \frac{d}{dx} \left( x \frac{\Delta L}{L} \right) = \frac{\Delta L}{L}$$
(79)

Thus for a uniformly stretched spring, y'(x) = dy(x)/dx is the amount of stretch per unit length, which we have called the *strain* of the spring.

If the strain is not uniform, if for example, we have a compressional wave on the spring, the strain is still given by

strain = 
$$\frac{local \ stretch}{per \ unit}$$
 =  $\frac{dy(x)}{dx}$  (80)

To help see that y'(x) = dy(x)/dx is the local amount of stretching per unit length, note that when we integrate the local stretching per unit length over the total length of the spring, we get the total stretch  $\Delta L$ .

$$\int_{0}^{L} \left(\frac{dy(x)}{dx}\right) dx = \int_{0}^{L} dy(x) = y(x) \Big|_{0}^{L}$$
$$= y(L) - y(0) \qquad (81)$$
$$= \Delta L - 0 = \Delta L$$

where  $y(L) = \Delta L$  the total displacement at the end.

Now go back to Equation (77)

$$\Delta T = kL\left(\frac{\Delta L}{L}\right)$$
 (77) repeated

which said that the change in tension in the spring is proportional to the strain  $\Delta L/L$ . We proved this was true for a uniform strain  $\Delta L/L$ . The obvious generalization when the strain is not uniform is to replace the average strain  $\Delta L/L$  by the local strain y'(x) = dy(x)/dx to get

$$\Delta T(x) = kL \frac{dy(x)}{dx} = kLy'(x)$$
(82)

where  $\Delta T(x)$  is the increase in the tension in a point x due to the local strain y'(x).

This gives us as the formula for the tension T(x) at point x

$$T(x) = T_0 + \Delta T(x)$$
  

$$T(x) = T_0 + kLy'(x)$$
(82a)

where  $T_0$  is the equilibrium tension, and kLy'(x) is the change in tension caused by the displacement of parts of the spring from their equilibrium position.

Let us now apply Equation (82a) to a short section of spring of length  $\Delta x$ , as shown in Figure (14). If the tension were uniform, the tension forces would cancel and there would be no net force on this section of the spring. A net force arises only if there is a change in tension as we go from x to  $x + \Delta x$ . This net force will be

$$net force = T(x + \Delta x) - T(x)$$

$$= \left[ T_0 + kLy'(x + \Delta x) \right] - \left[ T_0 + kLy'(x) \right]$$

$$= kL \left[ y'(x + \Delta x) - y'(x) \right]$$

$$= kL\Delta x \left[ \frac{y'(x + \Delta x) - y'(x)}{\Delta x} \right]$$
(83)

We immediately see that the last quantity in the square brackets is going to become, in the limit as  $\Delta x \rightarrow 0$ , the second derivative of y(x) with respect to x. Thus our formula for the net force on a section of length  $\Delta x$  is

$$\begin{array}{l}
\text{net} \\
\text{force} \\
\text{on } \Delta x
\end{array} = kL\Delta x \frac{d^2 y(x)}{dx^2} \tag{84}$$

If the spring has a mass per unit length of  $\mu$  kg/meter , the mass  $\Delta m$  of a length  $\Delta x$  is

$$\Delta m = \mu \Delta x \tag{85}$$

$$T(x) \xleftarrow{\Delta x \rightarrow} T(x + \Delta x)$$

Figure 14

There will be a net force on this short section of spring if the tension changes as we go from x to  $x+\Delta x$ .

If we allow waves on the spring, the displacement y(x) from equilibrium depends not only on the position x down the spring, but also on the time (t). Thus the displacement is described by the function y(x,t). The acceleration  $a_y(x,t)$  at position x on the spring is

$$a_{y}(x,t) = \frac{\partial^{2} y(x,t)}{\partial t^{2}}$$
(86)

We are using the partial derivative symbol  $\partial$  because we want to measure the change in y(x,t) with time at a fixed position x.

In terms of partial derivatives, Equation (84) for the net force on  $\Delta x$  is

With Equations (84a), (85) and (86), Newton's second law applied to  $\Delta m$  gives

$$\begin{array}{ll} net \\ force &= (\Delta m) \ a_y(x,t) & \begin{array}{c} Newton's \ law \\ F &= ma \end{array} \\ on \ \Delta m & \end{array}$$

$$kL\Delta x \frac{\partial^2 y(x,t)}{\partial x^2} = (\mu \Delta x) \frac{\partial^2 y(x,t)}{\partial t^2}$$
(87)

The factors of  $\Delta x$  cancel and we are left with

$$\left(\frac{kL}{\mu}\right)\frac{\partial^2 y(x,t)}{\partial x^2} = \frac{\partial^2 y(x,t)}{\partial t^2}$$
(88)

We recognize Equation (88) as the wave equation

$$v_{wave}^2 \frac{\partial^2 y(x,t)}{\partial x^2} = \frac{\partial^2 y(x,t)}{\partial t^2}$$

where we can identify the wave speed as

$$v \begin{pmatrix} compressional \\ Slinky wave \end{pmatrix} = \sqrt{\frac{kL}{\mu}}$$
 (89)

We got this same answer on page (15-8) of the Physics text using dimensional analysis. However, with dimensional analysis we were not sure whether a factor of 2 or  $\pi$  might be missing. Having derived the wave equation, we know that  $\sqrt{kL/\mu}$  is the correct answer with no missing constant factors.

# The Speed of Sound

The analysis of compressional sound waves in air can be carried out along lines very similar to our analysis of a compressional wave on a spring. However to do this, we need to build on our discussion of the behavior of an ideal gas in Chapters 17 and 18 of the Physics text. Thus we will assume that the reader is familiar with this material, including the discussion on adiabatic expansion in the Chapter 18 appendix.

Consider a column of gas with a cross sectional area A and length L as shown in Figure (15). We can think of the gas as being in a cylinder with frictionless walls, but it could be a hypothetical column in a large volume of gas. Let the variable x measure the distance down the column, starting at the left end, and imagine that we have a frictionless piston at the right end.

If we pull the piston out a small distance  $\Delta L$ , we change the volume of the gas by an amount

$$\Delta V = A\Delta L \tag{90}$$

and in so doing, decrease the pressure p.

How much the pressure changes depends upon the way the gas is expanded. If we expand it very slowly so that heat has time to flow into the gas and the temperature remains constant (this is called an *isothermal expansion*) then we have, from the ideal gas law

$$pV = NRT = constant$$
 *isothermal* (91)

where N is the number of moles of gas in the cylinder, R is the gas constant, and T the temperature in kelvins.

However in a sound wave, expansions and compressions happen so rapidly that there is not enough time for heat to flow in or out, and the temperature changes.



**Figure 15** Column of gas of cross-sectional area A and length L.

When we expand a gas with no heat flow, we call this an *adiabatic expansion*. As we saw in the appendix to Chapter 18, in an adiabatic expansion the gas obeys the equation

$$pV^{\gamma} = constant$$
  $adiabatic expansion$  (92)

where

$$\gamma = \frac{c_p}{c_v} \tag{93}$$

is the ratio of the specific heat  $c_p$  at constant pressure to the specific heat  $c_v$  at constant volume. It is Equation (92) for an adiabatic expansion rather than Equation (91) for an isothermal expansion that we need to use to describe the relationship between pressure and volume for a sound wave.

The quantity  $\gamma = c_p/c_v$  depends, as we saw at the beginning of Chapter 18, on the number of effective degrees of freedom of the gas molecules. As you found if you did Exercise 2 of Chapter 18, for a monatomic gas like helium or argon with no rotational degrees of freedom,  $\gamma = 1.66$  (5/3). For diatomic gases like oxygen, nitrogen, and of course air, that have two rotational degrees of freedom,  $\gamma = 1.40$ . When we get to more complex structures like CO<sub>2</sub> and NH<sub>4</sub>, then  $\gamma$  drops to 1.28.

We will now use Equation (92) for an adiabatic expansion to calculate the change  $\Delta p$  in pressure when we change the volume of the gas in the cylinder by an amount  $\Delta V$ . Before we compress we have

$$pV^{\gamma} = p_0 V_0^{\gamma} \tag{94}$$

where  $p_0$  and  $V_0$  are our original pressure and volume. After the expansion, V goes to  $V_0 + \Delta V$  and p goes to  $p_0 + \Delta p$ , where we know that  $\Delta p$  is negative for an expansion. Thus after the expansion we have

$$pV^{\gamma} = (p_0 + \Delta p)(V_0 + \Delta V)^{\gamma}$$
(95a)

With  $pV^{\gamma} = p_0 V_0^{\gamma}$  = constant, we get

$$p_0 V_0^{\gamma} = (p_0 + \Delta p)(V_0 + \Delta V)^{\gamma}$$
(95b)

We can use the fact that  $\Delta V$  is very small compared to  $V_0$  to get

$$(\mathbf{V}_0 + \Delta \mathbf{V})^{\gamma} = \left[\mathbf{V}_0 \left(1 + \frac{\Delta \mathbf{V}}{\mathbf{V}_0}\right)\right]^{\gamma} = \mathbf{V}_0^{\gamma} \left(1 + \frac{\Delta \mathbf{V}}{\mathbf{V}_0}\right)^{\gamma}$$

Using the approximation  $(1 + \alpha)^{\gamma} \approx 1 + \gamma \alpha$  for a small  $\alpha$ , we have

$$\left(1 + \frac{\Delta V}{V_0}\right)^{\gamma} \approx 1 + \gamma \frac{\Delta V}{V_0}$$
(96)

Using (96) in (95b), with  $p_0 + \Delta p = p_0(1 + \Delta p/p_0)$ , gives

$$p_0 V_0^{\gamma} = (p_0 + \Delta p)(V_0 + \Delta V)^{\gamma}$$

$$= p_0 \left(1 + \frac{\Delta p}{p_0}\right) V_0^{\gamma} \left(1 + \gamma \frac{\Delta V}{V_0}\right)$$
(97)

Multiplying this out gives

$$p_0 V_0^{\gamma} = p_0 V_0^{\gamma} \left[ 1 + \frac{\Delta p}{p_0} + \frac{\gamma \Delta V}{V_0} + \frac{\gamma}{p_0 V_0} \Delta p \Delta V \right]$$
(98)

The factors  $p_0 V_0^{\gamma}$  cancel, and we can neglect the second order term  $\Delta p \Delta V$ , giving

$$1 = 1 + \frac{\Delta p}{p_0} + \frac{\gamma \Delta V}{V_0}$$

After canceling the 1's and multiplying through by  $p_0$  we get for the pressure change  $\Delta p$ 

$$\Delta \mathbf{p} = -\gamma \mathbf{p}_0 \left( \frac{\Delta \mathbf{V}}{\mathbf{V}_0} \right) \tag{99}$$

If you look at the appendix to Chapter 18 in our discussion of the adiabatic expansion, you see that we started with the equation

$$\gamma p_0 \Delta V + \Delta p V_0 = 0 \tag{18-A8}$$

[which is Equation (99) if we solve for  $\Delta p$ ] and went through a number of calculus steps to derive  $pV^{\gamma} = constant$ . What we have done in going from  $pV^{\gamma} = constant$  to Equation (99) is to undo the calculus steps in that appendix. However one typically remembers the equation  $pV^{\gamma} = constant$  for adiabatic expansions rather than Equation (18-A8), and it seemed worthwhile to show how to get from  $pV^{\gamma} = constant$  to our formula for  $\Delta p$ .

Now that we have Equation (99) for  $\Delta p$ , we can follow essentially the same steps that we did earlier to calculate the speed of a compressional wave pulse on a spring.

If the cylinder in Figure (15) has a cross sectional area A, length L, and we move the piston out a distance  $\Delta L$ , we have

$$V_0 = AL$$
$$\Delta V = A\Delta L \tag{100}$$

thus from Equation (99) we have

$$\Delta p = -\gamma p_0 \left(\frac{\Delta V}{V}\right) = -\gamma p_0 \left(\frac{A\Delta L}{AL}\right)$$
$$= -\gamma p_0 \frac{\Delta L}{L}$$
(101)

In moving the piston out, the average displacement of a molecule y(x) at position x will be

$$y(x) = \frac{x}{L}\Delta L \tag{102}$$

which is the same as our Equation (78) for the average displacement of a piece of spring at position x. Differentiating Equation (102) with respect to x gives

$$y'(x) = \frac{dy(x)}{dx} = \frac{\Delta L}{L}$$
(103)

Thus we see that for a uniform displacement of the gas molecules, the strain, the displacement per unit length, is y'(x) = dy(x)/dx. We will now assume that even for non uniform displacements such as the kind we would have in a pressure pulse, y'(x) represents the local strain or displacement per unit length. In terms of this local strain, our formula (101) for  $\Delta p(x)$  becomes

$$\Delta p(x) = -\gamma p_0 y'(x) \qquad \begin{array}{c} local \\ pressure \\ change \end{array} \tag{104}$$

As in our discussion of springs, we can write this equation in the form

$$p(x) = p_0 + \Delta p(x)$$
  
 $p(x) = p_0 - \gamma p_0 y'(x)$ 
(105)

where we see that variations from the static pressure  $p_0$  are caused by local strains y'(x).

Now consider a section of the cylinder of length  $\Delta x$ located at x as shown in Figure (16). The gas external to  $\Delta x$  on the left, where the pressure is p(x), exerts a right directed force of magnitude

$$\mathbf{F}(\mathbf{x}) = \mathbf{A}\mathbf{p}(\mathbf{x}) \tag{106}$$

while the gas on the right exerts a left directed force of magnitude

$$F(x + \Delta x) = Ap(x + \Delta x)$$
(107)

where we have used the fact that the force is the pressure times the area. The net force on  $\Delta x$  is thus

$$F_{\text{net on }\Delta x} = F(x) - F(x + \Delta x)$$

$$= A \Big[ p(x) - p(x + \Delta x) \Big]$$
(108)

Using Equation (105) for p(x) we get

$$F_{net} = A\left\{ \left[ p_0 - \gamma p_0 y'(x) \right] - \left[ p_0 - \gamma p_0 y'(x + \Delta x) \right] \right\}$$

The  $p_0$  terms cancel and we are left with

$$\vec{F}_{net} = A\gamma p_0 \Big[ y'(x + \Delta x) - y'(x) \Big]$$

We can multiply by  $\Delta x / \Delta x$  to get

$$F_{\text{net}} = A\gamma p_0 \Delta x \left[ \frac{y'(x + \Delta x) - y'(x)}{\Delta x} \right]$$
(109)

As in the case of the spring, we will end up taking the limit as  $\Delta x$  goes to zero, so that the term in the square brackets in Equation (109) becomes the second derivative  $d^2y(x)/dx^2$ .

<∆x>			
area A	<u>p(x)</u> >		$\leq \frac{p(x+\Delta x)}{\Delta x}$
	>	(	

Figure 16

Pressure forces acting on a small section of gas in our hypothetical cylinder.

We will also let y(x) become a function of time y(x,t), so that second derivative becomes a partial derivative with respect to x only, and we get

$$F_{\text{net}} = \gamma p_0(A\Delta x) \frac{d^2 y(x,t)}{dx^2}$$
(110)

as our final formula for the net force on the gas in  $\Delta x$ .

The next step is to calculate the mass  $\Delta m$  of the gas in the region  $\Delta x$ . If the density of the gas is  $\rho$  kg/meter<sup>3</sup> and the volume inside  $\Delta x$  is  $(A\Delta x)$  meters<sup>3</sup>, we have

$$\Delta m = \rho A \Delta x \tag{111}$$

The acceleration of the gas in  $\Delta x$  is

$$a_{x}(t) = \frac{\partial^{2} y(x,t)}{\partial t^{2}}$$
(112)

Using Equations (110), (111), and (112) in Newton's second law gives

$$F_{\text{net on }\Delta x} = \Delta m a_{x}(t)$$
  

$$\gamma p_{0}(A\Delta x) \frac{\partial^{2} y(x,t)}{\partial x^{2}} = \rho(A\Delta x) \frac{\partial^{2} y(x,t)}{\partial t^{2}} \quad (113)$$

The factor  $A\Delta x$  cancels and we are left with

$$\frac{\gamma p_0}{\rho} \frac{\partial^2 y(\mathbf{x}, t)}{\partial \mathbf{x}^2} = \frac{\partial^2 y(\mathbf{x}, t)}{\partial t^2}$$
(114)

and we get the wave equation

$$v_{\text{wave}}^2 \frac{\partial^2 y(x,t)}{\partial x^2} = \frac{\partial^2 y(x,t)}{\partial t^2}$$
 (74) repeated

where we immediately see that the speed of the sound wave is given by

$$v_{sound} = \sqrt{\frac{\gamma p_0}{\rho}}$$
 (115)

In our discussion of sound waves in Chapter 15 of the Physics text, where we used dimensional analysis to predict the speed of sound, we came up with the formula

$$v_{\text{sound}} = \sqrt{\frac{B}{\rho}}$$
 (116)

where

$$\mathbf{B} \equiv \frac{\Delta \mathbf{p}}{\Delta \mathbf{V}/\mathbf{V}}$$

was called the *bulk modulus* of the gas. Going back to Equation (101), we have

$$\Delta p = = -\gamma p_0 \frac{\Delta L}{L} \tag{101}$$

$$= -\gamma p_0 \left(\frac{\Delta V}{V}\right) \tag{117}$$

for an adiabatic expansion, and the same with a + sign for compression. Thus

$$\frac{\Delta p}{\Delta V/V} = \gamma p_0 = B \qquad \begin{array}{c} for \ adiabatic \\ compression \end{array} \tag{118}$$

and our old formula for the speed of sound can be written as

$$v_{\text{sound}} = \sqrt{\frac{B}{\rho}} = \sqrt{\frac{\gamma p_0}{\rho}}$$
 (119)

which is the same result we got from the wave equation.

Using the ideal gas law, we can re-express the quantity  $p_0/\rho$  in our formula for the speed of sound in terms of the temperature T of the gas and some other constants. First we will write the density  $\rho$  as

$$\rho \frac{\text{kg}}{\text{meter}^3} = \frac{\text{M kg/mole} \times \text{N moles}}{\text{V meters}^3}$$
(120)

where M is the mass of one mole of the gas (an Avogadro's number of the gas molecules), N is the number of moles in our cylinder, and V the volume of the cylinder.

Next write the ideal gas law pV = NRT as

$$\frac{N}{V} = \frac{p}{RT}$$
(121)

where R is the gas constant and T the temperature in kelvins. Combining Equations (120) and (121) gives

$$\rho = M \frac{N}{V} = \frac{Mp}{RT}$$

or we have

$$\frac{p}{\rho} = \frac{RT}{M} \tag{122}$$

and our formula for the speed of sound becomes

$$v_{\text{sound}} = \sqrt{\frac{\gamma p}{\rho}} = \sqrt{\frac{\gamma RT}{M}}$$
 (123)

To interpret the physics of Equation (123), it is perhaps clearer to express the answer in terms of mass of the gas molecules involved. We have

$$m_{\text{molecule}} = \frac{M}{N_{\text{A}}} \frac{\text{kilograms/mole}}{\text{molecules/mole}}$$
$$= \frac{M}{N_{\text{A}}} \frac{\text{kilograms}}{\text{molecule}}$$

where N<sub>A</sub> is Avogadro's number, and

$$k = \frac{R}{N_A} \frac{joules/mole kelvin}{molecules/mole} = \frac{R}{N_A} \frac{joules}{kelvin}$$

is Boltzman's constant. Thus

$$\frac{R}{M} = \frac{N_A k}{N_A m_{\text{molecule}}} = \frac{k}{m_{\text{molecule}}}$$
(124)

and in terms of the molecular mass m<sub>molecule</sub> we get

$$v_{sound} = \sqrt{\frac{\gamma kT}{m_{molecule}}}$$
 (125)

From Equation (125), we immediately see that for a gas like hydrogen consisting of light molecules, the speed of sound is considerably greater than in a gas with heavy molecules.

### Exercise 8

Calculate the speed of sound at a temperature of 300 kelvin, in hydrogen, helium, nitrogen and  $CO_2$ . Use the fact that a hydrogen molecule has the mass of 2 protons, a helium atom the mass of 4 protons (with a nucleus of 2 protons and 2 neutrons), a nitrogen molecule the mass of 28 protons (each nucleus has 7 protons and usually 7 neutrons) and a  $CO_2$  molecule has a mass of around 44 protons (carbon nucleus has 6 protons and 6 or 7 neutrons, oxygen has 8 protons and 8 neutrons, for a total of 12 + 16 + 16 = 44 nuclear particles).

Aside from its dependence on the mass of the gas molecules, the other important feature is that the speed of sound is proportional to the square root of temperature. Thus the warmer the gas the greater the speed. This dependence of the speed of sound on the square root of temperature leads to a close connection between the speed of sound and the average speed of the air molecules due to their thermal motion.

In our discussion of the ideal gas law, we used the fact that the temperature was a measure of the average thermal kinetic energy of the gas, the precise relationship being

$$\frac{1}{2}m_{\text{molecule}}\bar{v}^2 = \frac{3}{2}kT \qquad (126)$$

where  $\bar{v}^2$  is the average of the square of the speed of the gas molecules  $(v^2 = v_x^2 + v_y^2 + v_z^2)$ . Writing Equation (126) in the form

$$\frac{kT}{m_{\text{molecule}}} = \frac{\bar{v}^2}{3} \tag{127}$$

and using this in Equation (125) gives

$$v_{\text{sound}} = \sqrt{\frac{\gamma kT}{m_{\text{molecule}}}} = \sqrt{\frac{\gamma \overline{v}^2}{3}}$$
$$v_{\text{sound}} = \overline{v} \sqrt{\frac{\gamma}{3}}$$
(128)

Several times we mentioned that the speed of sound is closely related to the speed of the air molecules due to their thermal motion. Equation (128) gives us the precise relationship. For air, for example, where  $\gamma = 1.28$  we get

$$v_{\text{sound}} = \bar{v} \sqrt{\frac{1.28}{3}} = .65 \bar{v}$$
 (129)

Sound travels over half as fast as the average speed  $\bar{v}$  of the air molecules.

# Calculus 2000-Chapter 3 The Gradient

The gradient operation represents the fundamental way that we go from a scalar field like the electric voltage V to a vector field like the electric field  $\vec{E}$ .

In this chapter, we present two distinct ways to introduce the gradient operation. One is to use the fact that electric fields are related to electric voltage the same way that forces are related to potential energy. The second, more geometrical way, is to picture the electric voltage as being described by a contour map, and that the electric field is described by the lines of steepest decent in the map. We present these two points of view as separate sections, View 1 and View 2, that can be read in either order.

We end the chapter with View 3, an application to fluids, where we see that the pressure force  $\vec{f}_p$  acting on fluid particles is the gradient of the pressure field p. This represents a straightforward example of obtaining a vector field  $\vec{f}_p$  from a scalar field p.

# **TWO VIEWS OF THE GRADIENT**

In the Physics text, our first laboratory exercise on electric phenomena was the potential plotting experiment illustrated in Figure (25-10) reproduced here. Two small brass cylinders connected to a battery were placed in a shallow tray of slightly conducting water. In order to measure the distribution of voltages V(x,y) at various points (x,y) in the water, we had two probes of bent, stiff, wire attached to blocks of wood, adjusted so that the tips of the wire stuck down in the water. The other end of the wire probes were attached to a voltmeter as shown.

By leaving one probe fixed, and moving the other in a way that the reading on the voltmeter remained constant, we could map out lines of constant voltage in the water. The results from a student lab notebook are shown slightly cleaned up in Figure (25-11). These lines of constant voltage are also known by the name *equipotential lines* or lines of equal electric potential. We also pointed out that these lines were analogous to lines of equal height, the contour lines in a contour map of the countryside.



#### Figure 25-10 (from Physics text)

Simple setup for plotting fields. You plot equipotentials by placing one probe (A) at a given position and moving the other (B) around. Whenever the voltage V on the voltmeter reads zero, the probes are at points of equal potential. While mapping the voltage V(x,y) at various points in the water was a straightforward process, our construction of the electric field lines  $\vec{E}(x,y)$  was not so obvious. Our procedure was to map  $\vec{E}(x,y)$  by drawing a set of lines perpendicular to the equipotential lines as shown in Figure (25-12). With this technique we were just barely able to tell whether the resulting field  $\vec{E}(x,y)$  more closely resembled the field of line charges or point charges. Our technique was conceptually correct, but a very crude way to determine the electric field  $\vec{E}(x,y)$  from a map of the voltage V(x,y).



#### Figure 25-11

Plot of the equipotential lines from a student project by B. J. Grattan. Instead of a tray of water, Grattan used a sheet of conductive paper, painting two circles with aluminum paint to replace the brass cylinders. We used the Adobe Illustrator® program to draw the lines through Grattan's data points.





After this initial experiment, we resorted to computer plots, like the one shown in Figure (25-15), to see the relationship between the electric field and a voltage map.

The computer plots, and the models we constructed from them, nicely illustrate the geometrical relationship between a voltage map and the electric field lines, but did not provide a convenient technique for actually calculating the field. The missing technique, which is the subject of this chapter, is the mathematical procedure called the *gradient*, a procedure involving the partial derivatives of the voltage function V(x,y).

As Figure (25-15) illustrates, there is a complete analogy between the contour map of a hilly terrain and electric field plots from a voltage map. We can

build our discussion of the gradient operation either upon our knowledge of the mathematics of the electric field, or by developing the ideas from a discussion of the nature of a hilly terrain. While both approaches are equivalent, we see the subject from two rather different points of view. The electric field approach is more efficient, while the hilly terrain approach develops some concepts that we will need later on.

As we mentioned in the introduction, we will begin this chapter with the electric field approach, and later discuss the hilly terrain viewpoint separately in **View 2**. You should study both approaches to see this important topic from two points of view. It does not really matter which one you study first.



Figure 25-15

Computer plot of the field lines and equipotentials for a charge distribution consisting of a positive charge + 3 and a negative charge - 1. These lines were then used to construct the plywood model.



# View 1 The Gradient from a Force – Energy Perspective

# CALCULATING THE ELECTRIC FIELD

Figure (1) shows a small section of the voltage map of Figure (25-15) on the previous page. The solid lines are equipotential lines, lines of constant voltage spaced .1 volts apart. We want to imagine that we actually have a detailed map of the voltage V(x,y) at every point (x,y), and want to mathematically determine, from that map, the electric field  $\vec{E}(x,y)$  at every point.

In the Physics text, we emphasized the idea that the electric voltage V(x,y) was the electric potential energy of a unit test charge, while the electric field  $\vec{E}(x,y)$  was the electric force on a unit test charge. Thus the connection between  $\vec{E}$  and V is the relationship between force and potential energy.

To review this relationship, imagine that I place a unit test particle at point A in Figure (1), where the voltage is  $V_A = .3$  volts. Since the voltage is the potential energy, in joules, of a unit test charge, our test particle at point A has a potential energy of .3 joules.



#### Figure 1

A small section of the voltage map, showing equipotential lines spaced .1 volts apart. We will calculate the amount of work required to move a unit test charge from point A to point B.

Now imagine that I move the test particle along the dashed line from point A at .3 volts over to point B at .4 volts. The potential energy of the particle has increased from .3 joules to .4 joules. Thus to move the particle, I must supply (.1) joules of energy to the particle.

Imagine that I move the test particle slowly, so that the force  $\vec{F}_{me}(x,y)$  that I exert on the particle is just enough to oppose the force  $\vec{E}(x,y)$  that the electric field is exerting on the particle. Thus for the entire trip from A to B we have

$$\vec{F}_{me}(x,y) = -\vec{E}(x,y) \tag{1}$$

The amount of work I do in moving the particle is given by the formula first discussed in Chapter 10 of the Physics text (see page 10-15, Equation (10-25)).

work I do in  
moving the  
test particle 
$$= \int_{A}^{B} \vec{F}_{me} \cdot d\vec{\ell}$$
(2)

Because I am moving the particle slowly so that all the work I do is stored as electric potential energy, and because the increase of potential energy of the unit test charge is  $V_B - V_A$ , we have

$$\int_{A}^{B} \vec{F}_{me} \cdot d\vec{\ell} = V_{B} - V_{A}$$
(3)

We can get *me* out of the equation by using Equation (1) to give

$$-\int_{A}^{B} \vec{E} \cdot d\vec{\ell} = V_{B} - V_{A}$$
(4)

Equation (4) is the integral equation that relates the voltage V(x,y) to the electric field  $\vec{E}(x,y)$ . It is a relationship we used extensively in the Physics text. In the Calculus text, we will often translate from integral to differential equations, and this chapter on the gradient will be our first example of how this is done.

The first step in going to a differential equation is to focus in on a very small region of Figure (1), a region shown in Figure (2), centered at the point  $(x_i,y_i)$  on the path from A to B. We have zoomed in so closely to the point  $(x_i,y_i)$  in Figure (2), we have so greatly magnified the plot, that the equipotential lines and the field lines in this region are simply straight lines at right angles to each other.

Now suppose we move our test particle from point (1) at  $(x_i, y_i)$  over a distance  $\overline{\Delta \ell}$  along the path to point (2) as shown. Equation (4) applied to this short displacement is

$$V_2 - V_1 = -\int_1^2 \vec{E}(x, y) \cdot d\vec{\ell}$$
 (5)

For this short path, we can assume that  $\vec{E}(x,y)$  is essentially constant and replace the integral by the product  $\vec{E}(x_i,y_i) \cdot \vec{\Delta \ell}$ , giving us

$$\mathbf{V}_2 - \mathbf{V}_1 = -\vec{\mathbf{E}}(\mathbf{x}_i, \mathbf{y}_i) \cdot \vec{\Delta \ell}$$
(6)

[You can see that in going from Equation (5) to (6) we are essentially undoing the step we took in Chapter (10) to derive the integral Equation (4).]

We are discussing the electric field of point charges. This is a conservative field, which is a fancy way of saying that the change in potential energy when we move a particle between two points does not depend upon the path we take. Thus if we first go a distance  $\Delta x$  along the x axis to point (3), then up the y axis a distance  $\Delta y$  to point (2), we should get the same change in voltage  $V_2 - V_1$  that we got by going directly from point (1) to point (2) along  $\overline{\Delta \ell}$ .



Figure 2

If we zoom in far enough, we reach a point where the equipotential lines and contour lines are straight lines perpendicular to each other.

In going along the x axis, we have

$$V_{3} - V_{1} = -\vec{E}(x_{i}, y_{i}) \cdot \vec{\Delta x}$$
  
=  $-E_{x}(x_{i}, y_{i}) \Delta x$  (7)

where the dot product of  $\vec{E}$  with the x directed displacement  $\Delta \vec{x}$  leaves us with the x component  $E_x$ . Writing out  $V_1$  and  $V_3$  in the form

$$V_1 = V(x_i, y_i)$$
$$V_3 = V(x_i + \Delta x, y_i)$$
Equation (7) becomes

$$V(x_i + \Delta x, y_i) - V(x_i, y_i) = -E_x(x_i, y_i)\Delta x \quad (8)$$

Dividing through by  $-\Delta x$  gives

$$E_{x}(x_{i},y_{i}) = -\frac{V(x_{i}+\Delta x,y_{i}) - V(x_{i},y_{i})}{\Delta x}$$

When we take the limit that  $\overrightarrow{\Delta \ell}$  goes to zero, both  $\overrightarrow{\Delta x}$  and  $\overrightarrow{\Delta y}$  will go to zero, giving

$$E_{x}(x_{i},y_{i}) = -\lim_{\Delta x \to 0} \left[ \frac{V(x_{i}+\Delta x,y_{i}) - V(x_{i},y_{i})}{\Delta x} \right]$$
(9)

By now you should recognize that the limit in Equation (9) is the partial x derivative of the function V(x,y) evaluated at the point  $(x_i,y_i)$ . Since this is true for any point (x,y), we get

$$E_{x}(x,y) = -\frac{\partial V(x,y)}{\partial x}$$
(10)

where the symbol  $\partial$  is used for partial derivatives.

#### **Exercise 1**

Use the above line of reasoning to show that

$$\mathsf{E}_{\mathsf{y}}(\mathsf{x},\mathsf{y}) = -\frac{\partial \mathsf{V}(\mathsf{x},\mathsf{y})}{\partial \mathsf{y}} \tag{11}$$

Introducing the unit vectors  $\hat{x}$  and  $\hat{y}$ , we can combine Equations (10) and (11) into the single vector equation

$$\vec{E}(x,y) = \hat{x} E_{x}(x,y) + \hat{y} E_{y}(x,y)$$
$$\vec{E}(x,y) = -\left[\hat{x} \frac{\partial V(x,y)}{\partial x} + \hat{y} \frac{\partial V(x,y)}{\partial y}\right]$$
(12)

Equation (12) is the differential equation we can use to calculate the electric field  $\vec{E}(x,y)$  at every point from a knowledge of the voltage V(x,y).

# Interpretation

$$\vec{E}(x,y) = -\left|\hat{x}\frac{\partial V(x,y)}{\partial x} + \hat{y}\frac{\partial V(x,y)}{\partial y}\right|$$
(12)

To help interpret Equation (12) repeated above, let us go back to Figure (25-15) where we started with a plot of the equipotential lines of the voltage V(x,y) and constructed a three dimensional plywood model of the voltage. The equipotential lines became the contour lines of this model, and the perpendicular electric field lines are the lines of steepest slope. If you were standing on terrain represented by this model, and the slope became slippery, the field line is the direction you would start to slide. Ski instructors call this direction of steepest slope the fall line.

To simplify the job of interpreting Equation (12), imagine that we are standing at the point  $A = (x_A, y_A)$  shown in Figure (3), where the contour line happens to be running in the y direction. If we move along a contour line there is no change in height, thus the partial derivative of V(x,y) with respect to y-the rate of change of V(x,y) in the y direction—is zero at point A.

$$\frac{\partial V(x,y)}{\partial y}\Big|_{\substack{x=x_A\\y=y_A}} = 0$$
(13)

The formula for  $\vec{E}(x,y)$  at point A becomes



Figure 3

The V = .2 volt contour line passes straight up through the point labeled A. Imagine that the surface is smoothed out and you walk along the dotted line.

To interpret Equation (14), imagine that we smooth out our plywood model of the voltage surface, then saw the model in two, cutting through the point A with the saw blade oriented along the x axis, along the dotted line in Figure (3). A side view of the upper piece is shown at the bottom of Figure (4). You can see that the voltage at the beginning of the cut, point C, is somewhat greater than .1 volts, and rises to just over .4 volts at the end, point D. The mathematical formula for the curve we see in Figure (4) is  $V(x, y_A)$ , and the partial derivative with respect to x at point A is the slope of the curve  $V(x,y_A)$  at  $x = x_A$ . This is just the tangent of the angle  $\theta$  in Figure (4).

$$\frac{\text{slope at point}}{\text{A going in}}_{\text{x direction}} \right\} = \frac{\partial V(x, y_A)}{\partial x} \bigg|_{x = x_A} = \tan \theta \quad (15)$$

This is the maximum slope at point A. If we sawed through point A, orienting the saw blade in any other direction, the slop at point A would be less. In particular the slope would be zero if we oriented the saw in the y direction.

From this discussion we see that the vector  $\vec{E}(x,y)$  points in the direction of the maximum slope and has a magnitude equal to that slope. The minus sign results from the fact that the force E is in the downward direction toward lower energy, while the positive slope, or *gradient* as we will call it, is in the upward direction.





S

χ

The top view shows the point A and the horizontal path through that point. The side view shows the path we would have to climb if the surface were smooth. The steepest slope at the point A is in the +x direction and is the tangent of the angle labeled  $\boldsymbol{\theta}$ .

# THE GRADIENT OPERATOR

The extension of Equation (12) to the case where the voltage varies in three dimensions, where V = V(x,y,z) is fairly obvious. It is

$$\vec{E}(x,y,z) = -\left[\hat{x} \frac{\partial V(x,y,z)}{\partial x} + \hat{y} \frac{\partial V(x,y,z)}{\partial y} + \hat{z} \frac{\partial V(x,y,z)}{\partial z}\right]$$
(16)

Until the beginning of the 20th century, research papers and textbooks dealing with partial derivatives used notation similar to Equation (16), and the formulas could become cumbersome and difficult to read. It was Willard Gibbs who introduced the *gradient operation*  $\vec{\nabla}$  defined by the equation

$$\vec{\nabla} \equiv \hat{x} \frac{\partial}{\partial x} + \hat{y} \frac{\partial}{\partial y} + \hat{z} \frac{\partial}{\partial z}$$
$$\equiv \hat{x} \nabla_{x} + \hat{y} \nabla_{y} + \hat{z} \nabla_{z}$$
(17)

where  $\nabla_{\mathbf{x}} = \partial/\partial \mathbf{x}$ , etc.

We call  $\vec{\nabla}$  an *operator* because it does not have an explicit meaning until it operates on something like the voltage function V(x,y,z).

$$\vec{\nabla} V(x,y,z) = \hat{x} \frac{\partial V}{\partial x} + \hat{y} \frac{\partial V}{\partial y} + \hat{z} \frac{\partial V}{\partial z}$$
  
=  $\hat{x} \nabla_x V + \hat{y} \nabla_y V + \hat{z} \nabla_z V$  (18)

With this notation, the formula for the electric field  $\vec{E}(x,y,z)$  in terms of the voltage V(x,y,z) is

$$\vec{E}(x,y,z) = -\vec{\nabla}V(x,y,z)$$
(19)

We say that the *electric field*  $\vec{E}$  *is minus the gradient of the voltage* V.

In the Physics text, we defined a vector field as a quantity with a vector value at every point in space. We began our discussion of vector fields in Chapter 23 with the velocity field rather than the electric field because the velocity field is easier to visualize. At any point in space the vector is simply the velocity vector of the fluid particle located there. For the electric field we first have to invent the concept of a tiny unit test charge before we can visualize the force vector at each point in space. Another mathematical concept, which we did not bother naming in the Physics text, is the *scalar field*. It is a quantity that has a scalar or numerical value at every point in space. An example of a scalar field is voltage, the potential energy of a unit test charge. At every point in space that we place the unit test charge, we get a voltage reading. Since energy has a magnitude but does not point anywhere, this reading has a *scalar* or numerical value only.

From Equation (19), we see that the gradient operator  $\vec{\nabla}$ , operating on a scalar field V creates the vector field  $\vec{E} = -\vec{\nabla}V$ . The vector  $\vec{\nabla}V$  has a numerical value equal to the maximum slope of V(x,y,z), and points opposite to the direction where the slope is greatest.

In the remainder of this part of the chapter, we will give examples of using the gradient operation to calculate the electric field from the voltage. In only a few cases, like the example of the parallel plate capacitor, is a Cartesian coordinate system (x,y,z)the most convenient coordinate system to use. In our study of electric and magnetic phenomena, we often dealt with point charges where there is spherical symmetry or line charges with cylindrical symmetry. We will see that to handle problems with spherical or cylindrical symmetry, it is much easier to work with the gradient  $\vec{\nabla} V$  expressed in spherical or cylindrical coordinate systems. Much of the detailed work for the remainder of the chapter will be to work out the formulas for the gradient in these coordinate systems. (You do these derivations once, and then use the results for the remainder of your scientific career.)

As we mentioned, we have View 2 later in the chapter, where we look at the gradient from a more geometrical and mathematical point of view. We end up with Equation (16) as the formula for the gradient, but explicitly demonstrate that the components  $\nabla_x V$  and  $\nabla_y V$  of the gradient transform (change) the same way the components of a displacement vector change when we go to a rotated coordinate system. Such discussions will become very useful later on.

# THE PARALLEL PLATE CAPACITOR

We introduced the parallel plate capacitor in Chapter 26, page 26-14 of the Physics text. We dealt with an idealized situation where we assumed that the plate diameters were much greater than the separation. Then we could neglect edge effects and assume that the electric field was uniform between the plates, as shown in Figure (26-27) reproduced here.

Since  $\vec{E}$  is the force on a unit test charge, and the voltage V is its potential energy, we can calculate the voltage V between the plates by calculating the amount of work required to lift the unit charge a distance y above the bottom plate. Since the force  $\vec{E}$  we have to work against is constant, the work we do is simply the force of magnitude E times the height y. If we say that the bottom plate is grounded, i.e., define the potential energy or voltage as being zero at the bottom plate, then the formula for the voltage between the plates is simply

$$V(x,y,z) = E y$$
 (20)

To evaluate E, we note that when we get to the top plate where y = d, the voltage is up to  $V_0$ , the voltage to which we charged the capacitor

$$V_0 = Ed \tag{21}$$

Thus  $E = V_0/d$ , and the voltage between the plates is given by

$$V(x,y,z) = \frac{V_0}{d} y$$
 (22)

Let us now turn the problem around and use the gradient formula  $\vec{E} = -\vec{\nabla}V$  to calculate the electric field  $\vec{E}$  from our voltage formula Equation (22). Writing out all the components of  $-\vec{\nabla}V$  as partial derivatives, we have from Equation (16)

$$\vec{\mathrm{E}}(\mathrm{x},\mathrm{y},\mathrm{z}) = -\left[\hat{\mathrm{x}}\frac{\partial \mathrm{V}}{\partial \mathrm{x}} + \hat{\mathrm{y}}\frac{\partial \mathrm{V}}{\partial \mathrm{y}} + \hat{\mathrm{z}}\frac{\partial \mathrm{V}}{\partial \mathrm{z}}\right]$$
(16)

The x partial derivative is

$$\frac{\partial V(x,y,z)}{\partial x} = \frac{\partial}{\partial x} \left( \frac{V_0 y}{d} \right) = 0$$
(23)

This is zero because there is no x dependence in our formula for V. When we take the partial derivative with respect to x, we hold y and z constant. Thus nothing in the formula  $V_0y/d$  changes when we change x, and this partial derivative is zero.

The other partial derivatives are

$$\frac{\partial V(x,y,z)}{\partial y} = \frac{\partial}{\partial y} \left[ \frac{V_0 y}{d} \right] = \frac{V_0}{d}$$
(24)

$$\frac{\partial V(x,y,z)}{\partial z} = \frac{\partial}{\partial z} \left[ \frac{V_0 y}{d} \right] = 0$$
(25)

Using Equations (23), (24), and (25) in (22) gives us

$$\vec{E} = -\hat{y}\frac{V_0}{d}$$
(26)

which says that  $\vec{E}$  points down in the  $-\hat{y}$  direction, and has a magnitude V<sub>0</sub>/d which we already know from Equation (21). We see that the calculation of  $\vec{E}$  from V using  $\vec{E} = -\vec{\nabla}V$  is a fairly straightforward process.



Figure 26-25

The parallel plate capacitor. The capacitor is charged up by connecting a battery across the plates as shown.



**Figure 26-26** *The electric field between and around the edge of the capacitor plates.* 



#### **Figure 26-27**

In our idealized parallel plate capacitor the field lines go straight from the positive to the negative plate, and the field is uniform between the plates.

# Voltage Inside a Conductor

The main idea of Chapter 26 of the Physics text was that you cannot have a static electric field inside a conductor if there is no flow of charge. The equivalent statement in terms of electric voltage is that the voltage is constant inside a conductor

$$V(x,y,z)_{inside a conductor} = constant$$
 (27)

To see that this gives a zero electric field, we have

$$\vec{E} = -\vec{\nabla} V_{\text{inside a conductor}} = 0$$
 (28)

All the components are zero because the partial derivative of a constant is zero.

To provide an explicit example, suppose we turn our parallel plate capacitor on its side and assume that it is constructed from thick metal plates as shown in Figure (5). The voltage as a function of distance is shown below the drawing of the plates. Inside the left plate the voltage has the constant value  $V_0$ , which gives zero field inside. Between the plates the voltage drops uniformly. It has a constant gradient, which gives us a constant electric field  $\vec{E} = -\vec{\nabla}V$  pointing in the direction of the downward slope. The voltage is again constant (V = 0) in the left hand plate.



#### Figure 5

Voltage in a parallel plate capacitor. The voltage is constant inside the plates and, for the assumed uniform field structure, drops uniformly between the plates.

# **ELECTRIC FIELD OF A POINT CHARGE**

Our first example of an electric field in the Physics text was the field of a point charge. If we have a charge Q located at the origin of our coordinate system, then the electric field at a position  $\vec{r} = (x,y,z)$  as shown in Figure (6) is given by

$$\vec{E}(r) = \hat{r} \frac{Q}{4\pi\epsilon_0 r^2} = \hat{r} \frac{kQ}{r^2}$$
 (29)

where  $\hat{r}$  is a unit vector in the  $\vec{r}$  direction and  $k=1/4\pi\epsilon_0$  .

In the Physics text, we mentioned, but never accurately derived, that the voltage V(r) of a point charge was

$$V(r) = \frac{Q}{4\pi\epsilon_0 r} = \frac{kQ}{r}$$
(30)

when we chose the zero of potential energy at r = infinity. What we want to do now is to show that the formula for  $\vec{E}(r)$  follows directly from Equation (30) for V(r) when we use the relationship

$$\vec{E} = -\vec{\nabla}V$$
 (14) repeated

The work is a bit messy, because we will be using a Cartesian coordinate system to solve a problem with spherical symmetry. Later we will find the formula for the gradient in spherical coordinates, and then see that it is very easy to evaluate  $\vec{E} = -\vec{\nabla}V$  for a point charge.



**Figure 6** *Out at a point given by the coordinate vector*  $\vec{r}$ , we have the unit vector  $\hat{r}$ .

Our first step will be to write out the vector equation  $\vec{E} = -\vec{\nabla}V$  as three component equations

$$E_x = -\frac{\partial V}{\partial x}$$
;  $E_y = -\frac{\partial V}{\partial y}$ ;  $E_z = -\frac{\partial V}{\partial z}$ 
(31)

Focusing on the x component equation we have

$$\mathbf{E}_{\mathbf{x}} = -\frac{\partial \mathbf{V}}{\partial \mathbf{x}} = -\frac{\partial}{\partial \mathbf{x}} \left( \frac{\mathbf{k}\mathbf{Q}}{\mathbf{r}} \right)$$

Taking the constant kQ outside the derivative we have

$$\mathbf{E}_{\mathbf{x}} = -\mathbf{k}\mathbf{Q}\,\frac{\partial}{\partial \mathbf{x}}\left(\frac{1}{\mathbf{r}}\right) \tag{32}$$

To go any farther, we have to express the distance r as a function of the coordinate x. This is done by the three dimensional Pythagorean theorem

$$r~=~\sqrt{x^2+y^2+z^2}$$

To calculate the derivative of (1/r) with respect to x now becomes an exercise in the use of the chain rule for differentiation. Let us start with

$$r^2 = x^2 + y^2 + z^2$$

which is easy to differentiate. We get

$$\frac{\partial r^2}{\partial x} = \frac{\partial}{\partial x} \left( x^2 + y^2 + z^2 \right) = 2x \tag{33}$$

Next look at

$$\frac{\partial \mathbf{r}}{\partial \mathbf{x}} = \frac{\partial}{\partial \mathbf{x}} \sqrt{\mathbf{r}^2} = \frac{\partial \sqrt{\mathbf{r}^2}}{\partial \mathbf{r}^2} \frac{\partial \mathbf{r}^2}{\partial \mathbf{x}}$$
(34)

To evaluate  $\partial \sqrt{r^2}/\partial r^2$ , set  $y = r^2$  so that we have, using  $\partial y^n/\partial y = ny^{n-1}$ 

$$\frac{\partial\sqrt{r^2}}{\partial r^2} = \frac{\partial\sqrt{y}}{\partial y} = \frac{\partial}{\partial y} y^{.5} = \frac{1}{2}y^{-.5} = \frac{1}{2r}$$
(35)

Thus using Equation (33) and (35) in (34) gives

$$\frac{\partial \mathbf{r}}{\partial \mathbf{x}} = \frac{\partial \sqrt{\mathbf{r}^2}}{\partial \mathbf{r}^2} \frac{\partial \mathbf{r}^2}{\partial \mathbf{x}^2} = \left(\frac{1}{2\mathbf{r}}\right) (2\mathbf{x})$$
$$\frac{\partial \mathbf{r}}{\partial \mathbf{x}} = \frac{\mathbf{x}}{\mathbf{r}}$$
(36)

which is a fairly simple result considering what we went through.

Finally we have

$$\frac{\partial}{\partial x} \left(\frac{1}{r}\right) = \frac{\partial r^{-1}}{\partial x} = \frac{\partial r^{-1}}{\partial r} \frac{\partial r}{\partial x}$$
$$= -r^{-2} \frac{\partial r}{\partial x} = -\frac{1}{r^{2}} \left(\frac{x}{r}\right)$$
$$\boxed{\frac{\partial}{\partial x} \left(\frac{1}{r}\right) = -\frac{x}{r^{3}}}$$
(37)

and our formula for  $E_x$  becomes

$$E_{x} = -kQ \frac{\partial}{\partial x} \left(\frac{1}{r}\right) = -kQ \left(-\frac{x}{r^{3}}\right)$$
$$E_{x} = kQ \frac{x}{r^{3}}$$
(38a)

Clearly the y and z components are

$$E_{y} = kQ \frac{y}{r^{3}}$$
(38b)

$$E_z = kQ \frac{z}{r^3}$$
(38c)

To check that we got the right answer, we can go back to Equation (29)

$$\vec{E}(r) = \hat{r} \frac{kQ}{r^2}$$
 (29) repeated

and replace the unit vector  $\hat{\mathbf{r}}$  with its definition  $\vec{\mathbf{r}}/r$  giving

$$\vec{r} = (r_x, r_y, r_z) = (x, y, z)$$

$$\hat{r} = \frac{\vec{r}}{r} = \frac{1}{r} (x, y, z)$$

$$\hat{r}_x = \frac{x}{r} ; \quad \hat{r}_y = \frac{y}{r} ; \quad \hat{r}_z = \frac{z}{r}$$
(39)

Equation (39) says, for example, that the x component of the unit vector  $\hat{r}$  has a length x/r. Thus the x component of  $\vec{E}$  in Equation (29) is

$$E_x = \hat{r}_x \frac{kQ}{r^2} = \frac{x}{r} \frac{kQ}{r^2} = kQ \frac{x}{r^3}$$
 (40)

with similar equations for  $E_y$  and  $E_z$ . Since Equations (38) and (40) are the same, we have verified that  $\vec{E} = -\vec{\nabla}V$  gives the correct result for V = kQ/r.

The messiness we encountered calculating the field of a point charge from V = kQ/r resulted from our calculating x, y, and z components of  $\vec{E}$  when we knew that  $\vec{E}$  pointed in the radial direction. If we use what is called a *spherical coordinate system*, we will find that the formula for the radial component of the electric field is simply

$$E_r = \frac{-\partial V(r)}{\partial r}$$
(41)

With V(r) = kQ/r we get

$$\mathbf{E}_{\mathbf{r}} = -\mathbf{k}\mathbf{Q}\,\frac{\partial}{\partial \mathbf{r}}\left(\frac{1}{\mathbf{r}}\right) = -\mathbf{k}\mathbf{Q}\left(-\frac{1}{\mathbf{r}^{2}}\right) = \frac{\mathbf{k}\mathbf{Q}}{\mathbf{r}^{2}} \quad (42)$$

and we get the final answer in a one line calculation.

To get this simple result requires, however, a fair amount of work deriving the formula for the gradient in spherical coordinates. First we have to define precisely what a spherical coordinate system is, show what the unit vectors are, and then calculate the components of the gradient when we move in the directions defined by the unit vectors. When this is all done, when we have the formula for the gradient in spherical coordinates, we can use the formula without ever going through the derivation again.

In the Physics text we encountered problems with plane symmetry, like the parallel plate capacitor, cylindrical symmetry, like the field of a line charge, and spherical symmetry like the field of a point charge we have just discussed. The plane symmetry problems are most easily handled in a Cartesian coordinate system, the cylindrical problems in what is called a *cylindrical coordinate system*, and spherical problems in a spherical coordinate system. We will now discuss these three coordinate systems and develop the formulas for the components of the gradient vector in each coordinate system. Since we have already done this for the Cartesian coordinate system, that discussion will serve as a review of the procedure we will use.

# GRADIENT IN THE CARTESIAN COORDINATE SYSTEM

An example of a right handed Cartesian coordinate system is shown in Figure (7). Out at some point  $\vec{r} = (x,y,z)$  the unit vectors  $\hat{x}$ ,  $\hat{y}$ , and  $\hat{z}$  are parallel to the x, y, and z axis as shown. It is called a right handed coordinate system because the unit vectors obey the relationship

$$\hat{\mathbf{x}} \times \hat{\mathbf{y}} = \hat{\mathbf{z}} \tag{43}$$

when we use the right hand rule for the cross product. (If we used a left hand rule, the z axis would point the other way.)

# **Exercise 1**

Show that  $\hat{y} \times \hat{z} = \hat{x}$  and  $\hat{z} \times \hat{x} = \hat{y}$ .

We will use the force/potential energy relationship to define the gradient vector. If I move a unit test charge a short distance  $\vec{\Delta \ell}$ , exerting a force  $\vec{F}_{me} = -\vec{E}$  to just overcome the electric field  $\vec{E}$ , the work  $\Delta W$  I do is

$$\Delta \mathbf{W} = \vec{\mathbf{F}}_{\mathrm{me}} \cdot \vec{\Delta \ell} = -\vec{\mathbf{E}} \cdot \vec{\Delta \ell}$$
(44)

Since this work is the change  $\Delta V$  in the potential energy of the unit test charge, we have

$$\Delta \mathbf{V} = -\vec{\mathbf{E}} \cdot \vec{\Delta \ell} \tag{45}$$

But the voltage V is related to the field  $\vec{E}$  by the gradient

 $\vec{E} = -\vec{\nabla}V$  (14) repeated





Using Equation (14) in (45), we can eliminate  $\vec{E}$  and get the relationship between the small change in voltage  $\Delta V$  and the voltage gradient  $\vec{\nabla} V$ 

$$\Delta \mathbf{V} = (\vec{\nabla}\mathbf{V}) \cdot \vec{\Delta \ell} \tag{46}$$

Equation (46) will allow us to find the formula for the gradient in the various coordinate systems.

To see how we are going to use Equation (46), we will start with the Cartesian coordinate system and choose  $\overrightarrow{\Delta \ell}$  to be a short step  $\Delta x$  in the x direction. Explicitly we will start at a point (x,y,z) and move to the point (x +  $\Delta x$ , y, z) so that  $\Delta V$ ,  $\overrightarrow{\Delta \ell}$  and  $(\overrightarrow{\nabla} V) \cdot \overrightarrow{\Delta \ell}$  become

$$\Delta V = V(x + \Delta x, y, z) - V(x, y, z)$$
(47)

$$\overline{\Delta \ell} = \hat{x} \Delta x \tag{48}$$

$$(\vec{\nabla}\mathbf{V})\cdot\vec{\Delta \ell} = (\vec{\nabla}\mathbf{V})_{\mathbf{X}}\,\Delta\mathbf{x} \tag{49}$$

Using (47) and (49) in (46) gives

$$V(x + \Delta x, y, z) - V(x, y, z) = (\vec{\nabla}V)_{x} \Delta x (50)$$

Dividing through by  $\Delta x$  and taking the limit as  $\Delta x$  goes to zero gives

$$\left(\vec{\nabla}V\right)_{x} = \lim_{\Delta x \to 0} \left[\frac{V(x + \Delta x, y, z) - V(x, y, z)}{\Delta x}\right]$$
(51)

which is the definition of the partial derivative. Thus

$$(\vec{\nabla}V)_{x} = \frac{\partial V(x, y, z)}{\partial x}$$
 (52)

which is our earlier result. This procedure does not give us anything new for a Cartesian coordinate system, but will give us new results for other coordinate systems.

(On the next page you will find two pictures of our model of the electric field of two point charges. We put the pictures there so that the discussion of the gradient in cylindrical and spherical coordinates would each be completed on facing pages.)



**Figure 25-14 (from Physics text)** Different views of the model of the electric field of two point charges  $Q_+ = +3$  and  $Q_- = -1$ .

# **GRADIENT IN CYLINDRICAL COORDINATES**

In a cylindrical coordinate system, we define the location of a point p by giving the distance  $\vec{r}$  out from the z axis, the angle  $\theta$  over from the x axis, and the height z above the xy plane as shown in Figure (8). The unit vectors are  $\hat{r}$  which points radially out from the z axis,  $\hat{z}$  which points in the z direction, and  $\hat{\theta}$  which is perpendicular to the  $\hat{r} \hat{z}$  plane. The direction of  $\hat{\theta}$  is the direction we move when increasing the angle  $\theta$ . This gives us a right handed coordinate system where the unit vectors are related by

$$\hat{\mathbf{r}} \times \hat{\mathbf{\theta}} = \hat{\mathbf{z}} \tag{53}$$

You should check for yourself that Equation (53) works for the unit vectors shown in Figure (8), and that  $\hat{\theta} \times \hat{z} = \hat{r}$  and  $\hat{z} \times \hat{r} = \hat{\theta}$ .

We will assume that in cylindrical coordinates, the gradient vector at point p is given by the equation

$$\vec{\nabla} V = \hat{r} \left( \vec{\nabla} V \right)_{r} + \hat{\theta} \left( \vec{\nabla} V \right)_{\theta} + \hat{z} \left( \vec{\nabla} V \right)_{z}$$
(54)

where  $(\vec{\nabla}V)_r$ ,  $(\vec{\nabla}V)_{\theta}$  and  $(\vec{\nabla}V)_z$  are the components of the gradient vector that we want to determine.



**Figure 8** The unit vectors  $\hat{r}$ ,  $\hat{\theta}$ ,  $\hat{z}$  in cylindrical coordinates.

To calculate the first component  $(\vec{\nabla}V)_r$ , we will start at the point p at  $(r, \theta, z)$  and move a short distance  $\Delta r$  in the  $\hat{r}$  direction, to the point  $(r + \Delta r, \theta, z)$ . Our change in voltage  $\Delta V$ , displacement  $\vec{\Delta} \vec{l}$  and the dot product  $(\vec{\nabla}V) \cdot \vec{\Delta} \vec{l}$  are for this move

$$\Delta V = V(r + \Delta r, \theta, z) - V(r, \theta, z)$$
 (55)

$$\vec{\Delta \ell} = \hat{\mathbf{r}} \,\Delta \mathbf{r} \tag{56}$$

$$(\vec{\nabla}\mathbf{V})\cdot\vec{\Delta \ell} = \left[\hat{\mathbf{r}}(\vec{\nabla}\mathbf{V})_{\mathbf{r}} + \hat{\mathbf{\theta}}(\vec{\nabla}\mathbf{V})_{\mathbf{\theta}} + \hat{\mathbf{z}}(\vec{\nabla}\mathbf{V})_{\mathbf{z}}\right]\cdot\hat{\mathbf{r}}\Delta\mathbf{r}$$
(57)

Since the unit vectors are all at right angles to each other,  $\hat{\mathbf{r}} \cdot \hat{\mathbf{r}} = 1$ ,  $\hat{\mathbf{\theta}} \cdot \hat{\mathbf{r}} = 0$  and  $\hat{\mathbf{z}} \cdot \hat{\mathbf{r}} = 0$ , giving us

$$\Delta \mathbf{V} = (\vec{\nabla}\mathbf{V}) \cdot \vec{\Delta \ell} = (\vec{\nabla}\mathbf{V})_{\mathbf{r}} \Delta \mathbf{r}$$
(58)

Dividing (58) through by  $\Delta r$ , using (55) for  $\Delta V$  and taking the limit as  $\Delta r$  goes to zero gives

$$(\vec{\nabla}V)_{r} = \lim_{\Delta r \to 0} \left[ \frac{V(r + \Delta r, \theta, z) - V(r, \theta, z)}{\Delta r} \right]$$
(59)

The right side of Equation (59) is what we will define to be the partial derivative of V(r,  $\theta$ , z) with respect to r in cylindrical coordinates

$$\frac{\partial V(r, \theta, z)}{\partial r} \equiv \lim_{\Delta r \to 0} \left[ \frac{V(r + \Delta r, \theta, z) - V(r, \theta, z)}{\Delta r} \right]$$
(60)

This is the rate of change of the function  $V(r, \theta, z)$  as we change the r coordinate. With this definition, we get

$$\left(\vec{\nabla}V\right)_{r} = \frac{\partial V(r, \theta, z)}{\partial r}$$
(61)

So far, our results look very much like what we had for Cartesian coordinates. However, we get something new when our step  $\overline{\Delta \ell}$  is in the  $\hat{\theta}$  direction. Suppose we are at the position  $(r, \theta, z)$ , and move to the new point  $(r, \theta + \Delta \theta, z)$  where we increased the  $\theta$  coordinate angle by  $\Delta \theta$  as shown in Figure (9). Since the angle  $\theta + \Delta \theta$  is measured in radians, the arc length  $\overline{\Delta \ell}$  that we move when going from  $\theta$  to  $\theta + \Delta \theta$  is

$$\Delta \ell = r \Delta \theta$$

You will notice that the vector displacement  $\overline{\Delta \ell}$  is in the same direction as the  $\hat{\theta}$  unit vector, thus

$$\Delta \hat{\ell} = \hat{\theta} r \Delta \theta \tag{62}$$

The change in voltage  $\Delta V$  and the dot product  $(\vec{\nabla} V) \cdot \vec{\Delta \ell}$  are thus

$$\Delta V = V(r, \theta + \Delta \theta, z) - V(r, \theta, z)$$
(63)

$$\vec{\nabla} \mathbf{V} \cdot \vec{\Delta \ell} = \left[ \hat{\mathbf{r}} (\vec{\nabla} \mathbf{V})_{\mathbf{r}} + \hat{\boldsymbol{\theta}} (\vec{\nabla} \mathbf{V})_{\boldsymbol{\theta}} + \hat{\mathbf{z}} (\vec{\nabla} \mathbf{V})_{\mathbf{z}} \right] \cdot \hat{\boldsymbol{\theta}} \mathbf{r} \Delta \boldsymbol{\theta}$$
$$= (\vec{\nabla} \mathbf{V})_{\boldsymbol{\theta}} \mathbf{r} \Delta \boldsymbol{\theta} \tag{64}$$

where we used  $\hat{\theta} \cdot \hat{\theta} = 1$ ,  $\hat{r} \cdot \hat{\theta} = \hat{z} \cdot \hat{\theta} = 0$ .

Using (63) and (64) in our equation  $\Delta V = \vec{\nabla} V \cdot \vec{\Delta \ell}$ , we get

$$V(r, \theta + \Delta \theta, z) - V(r, \theta, z) = (\vec{\nabla} V)_{\theta} r \Delta \theta \quad (65)$$



**Figure 9** The displacement  $\overline{\Delta \ell}$  when we increase the angle  $\theta$  by  $\nabla \theta$ .

Dividing Equation (65) through by  $r\Delta\theta$  and then taking the limit as  $\Delta\theta$  goes to zero gives

$$(\vec{\nabla}V)_{\theta} = \frac{1}{r} \left\{ \lim_{\Delta\theta \to 0} \left[ \frac{V(r, \theta + \Delta\theta, z) - V(r, \theta, z)}{\Delta\theta} \right] \right\}$$
(66)

We define the quantity in curly brackets to be the partial derivative of V(r,  $\theta$ , z) with respect to the variable  $\theta$ 

$$\frac{\partial V(r, \theta, z)}{\partial \theta} \equiv \lim_{\Delta \theta \to 0} \left[ \frac{V(r, \theta + \Delta \theta, z) - V(r, \theta, z)}{\Delta \theta} \right]$$
(67)

Thus we end up with the equation

$$(\vec{\nabla} \mathbf{V})_{\theta} = \frac{1}{r} \frac{\partial \mathbf{V}(r, \theta, z)}{\partial \theta}$$
(68)

and we get a factor of 1/r in our formula for the  $\theta$  component of the gradient in cylindrical coordinates. The factor of 1/r appears because the partial derivative with respect to  $\theta$  measures the rate of change of V for a given change  $\Delta\theta$  in angle, while the gradient measures the rate of change of V with respect to a given step in distance. When we make a change  $\Delta\theta$  in angle, the distance we move is  $r\Delta\theta$  which increases with r. The factor of r has to be divided out to get the rate of change of V with distance.

### Exercise 2

Following the above steps, show that

$$(\vec{\nabla} V)_{z} = \frac{\partial V(r, \theta, z)}{\partial z}$$
(69)

This should look the same as our derivation for the Cartesian coordinate system.

# **GRADIENT IN SPHERICAL COORDINATES**

While the steps are fresh, let us derive the formulas for the components of the gradient vector in spherical coordinates. We will then return to various applications of the new gradient formulas.

In the spherical coordinate system shown in Figure (10), a point p is located by the displacement  $\vec{r}$  from the origin, the angle  $\theta$  that the coordinate vector  $\vec{r}$  makes with the z axis, and the angle  $\phi$  that the projection of  $\vec{r}$  on the x,y plane makes with the x axis. The unit vectors are  $\hat{r}$  pointing out in the  $\vec{r}$  direction,  $\hat{\theta}$  which lies in the  $\hat{r}$   $\hat{z}$  plane pointing in the direction of increasing  $\theta$ , and  $\hat{\phi}$  which is perpendicular to the  $\hat{r}$   $\hat{z}$  plane, in the direction of increasing  $\phi$ . This gives us a right handed coordinate system where

$$\hat{\mathbf{r}} \times \boldsymbol{\theta} = \hat{\boldsymbol{\phi}} \tag{70}$$

(Again, show for yourself that  $\hat{\theta} \times \hat{\phi} = \hat{r}$  and  $\hat{\phi} \times \hat{r} = \hat{\theta}$ .)



**Figure 10** The unit vectors  $\hat{r}$ ,  $\hat{\theta}$ ,  $\hat{\phi}$  for a spherical coordinate system.



**Figure 11** *The step*  $\overline{\Delta \ell}$  *when we increase*  $\theta$  *by*  $\Delta \theta$ . *We are directly facing the rz plane.* 

#### Exercise 3

Start at the point  $(r, \theta, \phi)$  and move a distance  $\overrightarrow{\Delta \ell}$  to the point V(r +  $\Delta r, \theta, \phi$ ) and show that the r component of the gradient in spherical coordinates is

$$(\vec{\nabla}V)_{r} = \frac{\partial V(r, \theta, \phi)}{\partial r}$$
(71)

where

$$\frac{\partial V(r, \theta, \phi)}{\partial r} = \lim_{\Delta r \to 0} \left[ \frac{V(r + \Delta r, \theta, \phi) - V(r, \theta, \phi)}{\Delta r} \right]$$
(72)

It was Equation (71) that we used to show in one line that the voltage V = kQ/r leads to the field  $\vec{E} = \hat{r} kQ/r^2$ .

In spherical coordinates, the radial component of the gradient is simply the partial derivative, as we asked you to show in Exercise 3. We get new results when we look at the  $\theta$  and  $\phi$  components, where the change in distance  $\Delta \ell$  is not equal to  $\Delta \theta$  or  $\Delta \phi$  alone.

First let  $\overline{\Delta \ell}$  be in the  $\hat{\theta}$  direction, so that we go from the point (r,  $\theta$ ,  $\phi$ ) to (r,  $\theta + \Delta \theta$ ,  $\phi$ ). The distance  $\overline{\Delta \ell}$ is shown in Figure (11) where we are looking squarely at the rz plane. You can see that  $\overline{\Delta \ell}$  is in the  $\hat{\theta}$ direction and has a magnitude  $\Delta \ell = r\Delta \theta$  so that

$$\overline{\Delta \ell} = \hat{\theta} r \Delta \theta \tag{73}$$

The change in voltage  $\Delta V$  and the dot product  $\vec{\nabla} V \cdot \vec{\Delta \ell}$  are

$$\Delta V = V(r, \theta + \Delta \theta, \phi) - V(r, \theta, \phi)$$
(74)

$$\vec{\nabla} \mathbf{V} \cdot \vec{\Delta \ell} = \left[ \hat{\mathbf{r}} (\vec{\nabla} \mathbf{V})_{\mathbf{r}} + \hat{\boldsymbol{\theta}} (\vec{\nabla} \mathbf{V})_{\boldsymbol{\theta}} + \hat{\boldsymbol{\phi}} (\vec{\nabla} \mathbf{V})_{\boldsymbol{\phi}} \right] \cdot \hat{\boldsymbol{\theta}} \mathbf{r} \Delta \boldsymbol{\theta}$$
$$= (\vec{\nabla} \mathbf{V})_{\boldsymbol{\theta}} \mathbf{r} \Delta \boldsymbol{\theta} \tag{75}$$

where  $\hat{\theta} \cdot \hat{\theta} = 1$ ,  $\hat{r} \cdot \hat{\theta} = \hat{\phi} \cdot \hat{\theta} = 0$ .

Equating  $\Delta V$  from (74) with  $\vec{\nabla} V \cdot \vec{\Delta \ell}$  in (75), then dividing through by  $r\Delta \theta$  and taking the limit as  $\Delta \theta$  goes to zero, gives

$$(\vec{\nabla} \mathbf{V})_{\theta} = \frac{1}{r} \lim_{\Delta \theta \to 0} \left[ \frac{\mathbf{V}(\mathbf{r}, \theta + \Delta \theta, \phi) - \mathbf{V}(\mathbf{r}, \theta, \phi)}{\Delta \theta} \right]$$
(76)

We define the partial derivative of  $V(r, \theta, \phi)$  with respect to  $\theta$  in spherical coordinates as

$$\frac{\partial V(r, \theta, \phi)}{\partial \theta} \equiv \lim_{\Delta \theta \to 0} \left[ \frac{V(r, \theta + \Delta \theta, \phi) - V(r, \theta, \phi)}{\Delta \theta} \right]$$
(77)

so that we get

$$(\vec{\nabla}V)_{\theta} = \frac{1}{r} \frac{\partial V(r, \theta, \phi)}{\partial \theta}$$
(78)

as the formula for the  $\theta$  component of the gradient vector in spherical coordinates.

Finally we will derive the  $\phi$  component of  $\nabla V$  by taking a step  $\Delta \ell$  in the  $\phi$  direction. The geometry is shown in Figure (12). The first thing to note is that the projection of the coordinate vector  $\vec{r}$  down on the xy plane has a length ( $r \sin \theta$ ). This is the distance the point p is out from the z axis. When we rotate an angle  $\Delta \phi$  about the z axis, the arc length  $\Delta \ell$ out a distance ( $r \sin \theta$ ) is ( $r \sin \theta$ ) $\Delta \phi$ . This distance is in the direction of the unit vector  $\hat{\phi}$ , thus

$$\overline{\Delta \ell} = \hat{\phi} (r \sin \theta) \Delta \phi \tag{79}$$

The change in voltage, going from  $(r, \theta, \phi)$  to  $(r, \theta, \phi + \Delta \phi)$  is

$$\Delta V = V(r, \theta, \phi + \Delta \phi) - V(r, \theta, \phi)$$
(80)



#### Figure 12

\_\_\_\_

The step  $\overline{\Delta l}$  when we increase  $\phi$  by  $\Delta \phi$ . Note that we are out a distance  $r \sin \theta$ from the z axis.

The quantity  $\vec{\nabla} \mathbf{V} \cdot \vec{\Delta \ell}$  is

$$\vec{\nabla} \mathbf{V} \cdot \vec{\Delta} \boldsymbol{\ell} = \left[ \hat{\mathbf{r}} (\vec{\nabla} \mathbf{V})_{\mathbf{r}} + \hat{\boldsymbol{\theta}} (\vec{\nabla} \mathbf{V})_{\boldsymbol{\theta}} + \hat{\boldsymbol{\phi}} (\vec{\nabla} \mathbf{V})_{\boldsymbol{\phi}} \right] \cdot \hat{\boldsymbol{\phi}} (\mathbf{r} \sin \boldsymbol{\theta}) \Delta \boldsymbol{\phi}$$
$$= (\vec{\nabla} \mathbf{V})_{\boldsymbol{\phi}} (\mathbf{r} \sin \boldsymbol{\theta}) \Delta \boldsymbol{\phi}$$

(81)

because 
$$\hat{\phi} \cdot \hat{\phi} = 1$$
 and  $\hat{r} \cdot \hat{\phi} = \hat{\theta} \cdot \hat{\phi} = 0$ .

Equating  $\vec{\nabla} \mathbf{V} \cdot \vec{\Delta \ell}$  in Equation (81) to  $\Delta \mathbf{V}$  in (80) gives

$$V(r, \theta, \phi + \Delta \phi) - V(r, \theta, \phi) = (\overline{\nabla} V)_{\phi} (r \sin \theta) \Delta \phi$$
(82)

Dividing (82) through by  $(r \sin \theta)\Delta\phi$  and taking the limit at  $\Delta\phi$  goes to zero gives

$$\left(\vec{\nabla}V\right)_{\phi} = \frac{1}{r\sin\theta} \lim_{\Delta\phi\to 0} \left[\frac{V(r,\theta,\phi+\Delta\phi) - V(r,\theta,\phi)}{\Delta\phi}\right]$$
(83)

We define the partial derivative with respect to  $\phi$  in spherical coordinates as

$$\frac{\partial V(r, \theta, \phi)}{\partial \phi} = \lim_{\Delta \phi \to 0} \left[ \frac{V(r, \theta, \phi + \Delta \phi) - V(r, \theta, \phi)}{\Delta \phi} \right]$$
(84)

to get the result

$$\left(\vec{\nabla}V\right)_{\phi} = \frac{1}{r\sin\theta} \frac{\partial V(r,\theta,\phi)}{\partial\phi}$$
(85)
#### SUMMARY OF GRADIENT FORMULAS

We collect in one place the formulas for the gradient in Cartesian, cylindrical and spherical coordinates.

#### **Cartesian Coordinates**



#### **Cylindrical Coordinates**



**Spherical Coordinates** 



# **EXAMPLES**

# **Electric Field of a Point Charge**

Let us now see explicitly how the formula for the gradient in spherical coordinates, Equation (88), makes it easy to calculate the electric field of a point charge, starting from the voltage formula

$$V(r) = \frac{kQ}{r}$$
 (27) repeated

The formula for the gradient in spherical coordinates is

$$\vec{\nabla}$$
V =  $\hat{r}\frac{\partial V}{\partial r} + \frac{\hat{\theta}}{r}\frac{\partial V}{\partial \theta} + \frac{\hat{\phi}}{r\sin\theta}\frac{\partial V}{\partial \phi}$  (88) repeat

While Equation (88) looks somewhat messy, the thing to note is that V(r) has no dependence on the variables  $\theta$  and  $\phi$ , thus the partial derivatives with respect to these variables are zero

$$\frac{\partial V(r)}{\partial \theta} = 0 \quad ; \quad \frac{\partial V(r)}{\partial \phi} = 0 \tag{89}$$

and all we are left with is

$$\vec{\nabla} \mathbf{V} = \hat{\mathbf{r}} \frac{\partial \mathbf{V}(\mathbf{r})}{\partial \mathbf{r}} \tag{90}$$

We have for  $\partial V(r)/\partial r$ 

$$\frac{\partial}{\partial r} \left( \frac{kQ}{r} \right) = kQ \frac{\partial}{\partial r} (r^{-1}) = -1 \frac{kQ}{r^2}$$
(91)

thus we get

$$\vec{\mathsf{E}} = -\vec{\nabla}\mathbf{V} = -\left(-\hat{\mathbf{r}}\frac{\mathbf{k}\mathbf{Q}}{\mathbf{r}^2}\right) = \hat{\mathbf{r}}\frac{\mathbf{k}\mathbf{Q}}{\mathbf{r}^2} \tag{92}$$

which is the correct answer.

The advantage of using spherical coordinates to calculate the field of a point charge was that, two out of three of the components of the gradient were zero, and we had only a simple derivative for the remaining component. This is the kind of simplification you get when you use a coordinate system that matches the symmetry of the problem at hand. Our next example will be the calculation of the electric field of a line charge. That problem has cylindrical symmetry, and is most easily handled using a cylindrical coordinate system.

#### **Electric Field of a Line Charge**

In the Physics text, our first calculation of the electric field of an extended object was to show that the radially directed electric field of a charged wire, shown in Figure (24-27) repeated here, had a magnitude

$$E(r) = \frac{\lambda}{2\pi\epsilon_0 r}$$
 (24-43) repeated

where  $\lambda$  is the amount of charge per meter on the wire and r is the radial distance out from the wire. To simplify the constants, we will set  $k = 1/2\pi\epsilon_0$  so that the vector formula for  $\vec{E}$  is

$$\vec{E}(r) = \hat{r} \frac{k\lambda}{r}$$
;  $k = \frac{1}{2\pi\epsilon_0}$  (93)

In the Physics text we never did say what the voltage was in the vicinity of a charged wire. You will see why shortly.



#### Figure 24-27 (repeated)

Using Gauss' law to calculate the electric field of a line charge. Draw the Gaussian surface around a section of the rod. The flux all flows out through the cylindrical surface. We can assume, because of the cylindrical symmetry of the problem, that the voltage V depends only on the radial distance r out from the wire. That is, that V = V(r). Thus the partial derivatives with respect to the variables  $\theta$  and z (using cylindrical coordinates) should be zero and we should be left with

$$\vec{\mathbf{E}} = -\vec{\nabla}\mathbf{V} = -\left(\hat{\mathbf{r}}\frac{\partial\mathbf{V}(\mathbf{r})}{\partial\mathbf{r}} + \frac{\hat{\theta}}{\mathbf{r}}\frac{\partial\mathbf{V}(\mathbf{r})}{\partial\theta} + \hat{\mathbf{z}}\frac{\partial\mathbf{V}(\mathbf{r})}{\partial\mathbf{z}}\right)$$
$$= -\hat{\mathbf{r}}\frac{\partial\mathbf{V}(\mathbf{r})}{\partial\mathbf{r}} \qquad (94)$$

where we used Equation (87) for the gradient in cylindrical coordinates.

Comparing Equations (93) and (94) for  $\vec{E}$  we get

$$\vec{E} = \hat{r} \frac{k\lambda}{r} = \hat{r} \left( -\frac{\partial V(r)}{\partial r} \right)$$
 (95)

As a result, the voltage V(r) should obey the equation

$$\frac{\partial \mathbf{V}(\mathbf{r})}{\partial \mathbf{r}} = -\mathbf{k}\lambda\left(\frac{1}{\mathbf{r}}\right) \tag{96}$$

The question we have now is, what function of r, when differentiated with respect to r, gives 1/r? The answer, you may recall from Chapter 1 of the Calculus text, is the *natural logarithm*. Explicitly

$$\frac{\mathrm{d}}{\mathrm{d}r}(\ln r) = \frac{1}{r} \tag{97}$$

Thus the appropriate voltage V(r) is

$$V(r) = -k\lambda \ln r$$
(98)

Going back from this V(r) to  $\vec{E}$  we have

$$\vec{\nabla} \mathbf{V}(\mathbf{r}) = \hat{\mathbf{r}} \frac{\partial}{\partial \mathbf{r}} (-\mathbf{k} \,\lambda \ln \mathbf{r}) = \hat{\mathbf{r}} \left( -\mathbf{k} \,\lambda \frac{\partial \ln \mathbf{r}}{\partial \mathbf{r}} \right) = \hat{\mathbf{r}} \left( \frac{-\mathbf{k} \,\lambda}{\mathbf{r}} \right)$$
(99)

and

$$\vec{\mathbf{E}}(\mathbf{r}) = -\vec{\nabla}\mathbf{V}(\mathbf{r}) = \hat{\mathbf{r}}\left(\frac{\mathbf{k}\lambda}{\mathbf{r}}\right)$$
 (100)

This explicitly checks that the voltage  $-(k\lambda \ln r)$  leads to the electric field of a line charge.

The logarithm ln(r) that appears in Equation (100) is an interesting function in that it is zero at r = 1, goes to  $-\infty$  at r = 0 and  $+\infty$  at  $r = \infty$  as shown in Figure (13). Thus, for example, at r = 0 we get

$$V(r) \Big|_{r=0} = -kr \ln(0) = -kr (-\infty) = +\infty \quad (101)$$

and the voltage becomes infinite. This tells us that it is not physically reasonable to put a finite charge density  $\lambda$  on an infinitely thin wire. We had the same problem with a point charge. The formula V = kQ/r also goes to infinity at r = 0 which tells us we have a problem with the potential energy of a point charge of zero radius. (The modern theory of quantum electrodynamics treats the electron as a point charge of zero radius. The tricky part of the theory is to get around the infinities that result from this.)



Figure 13

The function ln(x) starts out at minus infinity at x = 0, goes through zero at x = 1, and slowly goes to plus infinity at x = infinity.

At large distances, there is no problem with the formula for the voltage of a *point charge*. At  $r = \infty$ , the voltage V = kQ/r goes to zero, which is what we wanted for the potential energy of a test charge infinitely far away. But for a *line charge*, Equation (94) gives

$$V(\mathbf{r})\Big|_{\mathbf{r}\,=\,+\,\infty} = -\,\mathbf{k}\,\lambda\,\mathbf{ln}(+\,\infty) \tag{102}$$

This predicts a voltage or potential energy of *minus infinity* when we are infinitely far away from a line charge! How did this happen?

Either the mathematics is wrong, or our physical interpretation is wrong. The answer lies with the physical interpretation. What is wrong is that *you cannot get infinitely far away from a line charge*. Any real physical piece of wire must have a finite length. The wire may look infinitely long when you are close to it, but as you move away, you will eventually be able to see both ends. The farther away you move, the shorter the wire looks. Move infinitely far from the wire and the wire looks like a point charge and the voltage it produces goes to zero. Thus physically we will not encounter the infinity that appears at large distances in the formula for the voltage of a line charge.

As we have often mentioned, in any formula for potential energy, we can arbitrarily choose the zero of potential energy (the floor) wherever we want. For point charges, we usually choose the zero of potential energy out at  $r = \infty$ . We have seen that we cannot make the same choice for a line charge. What we have to do is write the formula for the potential energy in the more general form

$$V(r) = -k \lambda \ln(r) + constant$$
 (98a)

and adjust the constant so that V(r) is zero at some convenient place. We can see how this works in the following discussion of a coaxial cable.

#### The Coaxial Cable

A physical example where our voltage formula (98a) makes sense is the coaxial cable. Suppose we have a cable whose inner conductor has a radius  $r_i$  and the outer shield has an inside radius  $r_0$  as shown in Figure (14). Assume that the inner conductor has a charge density  $\lambda$  coulombs per meter, and the outer conductor is grounded (i.e., we say that the voltage V(r) is zero at  $r = r_0$ .) What is the voltage throughout the cable?

First of all, we know that the voltage inside a conductor must be constant so that the field  $\vec{E} = -\vec{\nabla}V$  inside is zero. Since the outer conductor is grounded, the voltage throughout the shield (for  $r > r_0$ ) will be zero as shown in Figure (15). The voltage on the inner conductor will have some constant value  $V_i$  (for  $r < r_0$ ).

Between the conductors, in the region between  $r_i$  and  $r_0$ , the voltage must have the logarithmic dependence given by Equation (98a)

$$V(\mathbf{r}) = -\mathbf{k}\,\lambda\,\ln\mathbf{r} + \text{constant} \tag{103}$$

We can evaluate the constant by setting the voltage equal to zero out at the grounded shield, at  $r = r_0$ . This gives

 $V(r_0) = -k \lambda \ln r_0 + \text{constant} = 0$ constant = k \lambda \lnr\_0 (104)

and V(r) becomes

$$V(\mathbf{r}) = -\mathbf{k}\,\lambda\,\mathbf{lnr} + \mathbf{k}\,\lambda\,\mathbf{lnr}_0 \tag{105}$$

Logarithms have the peculiar property



#### Figure 14

A coaxial cable, where the inner wire has a radius  $r_i$ and the outer grounded shield an inner radius  $r_0$ .

Thus V(r) in Equation (99) can be more compactly written

$$\mathbf{V}(\mathbf{r}) = \mathbf{k} \,\lambda \ln \left(\frac{\mathbf{r}_0}{\mathbf{r}}\right) \tag{107}$$

With the constant k written out as  $1/2\pi\varepsilon_0$  (see Equation 93), we get

$$\mathbf{V}(\mathbf{r}) = \frac{\lambda}{2\pi \varepsilon_0} \ln\left(\frac{\mathbf{r}_0}{\mathbf{r}}\right)$$
(108)

At the outer shield, at  $r = r_0$ , we have

 $\ln(r_0/r) = \ln(1) = 0$ 

and the voltage goes to zero. This is what we wanted for a grounded shield.

As demonstrated in Exercise 4 below, Equation (108) allows us to calculate the charge density  $\lambda$  on the inner conductor of a coaxial cable when the outer conductor is grounded and the inner conductor is raised to some voltage  $V_i$ .

#### Exercise 4

(a) For the coaxial cable of Figure (14), find the formula for the charge density  $\lambda$  when the inner conductor is at a voltage V<sub>i</sub> volts.

(b) Suppose V<sub>i</sub> = 100 volts,  $r_i$  = .5 mm,  $r_0$  = 2 mm and recall that  $\epsilon_0 \approx 9 \times 10^{-12}$ . Then what is  $\lambda$  in coulombs per meter?

(c) What is the general formula for the capacitance per meter of the coaxial cable in Figure (14)?



**Figure 15** *Voltage in the coaxial cable.* 

# View 2 The Gradient from a Geometrical Perspective

In the first part of this chapter, we used the relationship between force and potential energy to define what we meant by the gradient vector. We then used that relationship to derive the formulas for the gradient in cylindrical and spherical coordinates.

What we want to do now is to approach the gradient from a geometrical point of view. This is the point of view we began to develop when we constructed the physical models of electric voltage like the one shown in Figure (25-15) reproduced again here. Once we have developed a geometrical definition of the gradient we will check that the gradient behaves like a vector. To do that, we show that the components of the gradient change or transform the same way that the components of a displacement vector when we rotate the coordinate system. This idea of testing the vector nature of a new quantity will become particularly important when we get to a mathematically advanced discussion of special relativity.

This discussion of the gradient is designed to be independent of the first part of the chapter, so that you can start from either approach. This leads to some repetition of definitions, but the points of view are sufficiently different that some duplication should not be a problem. We, of course, end up with the same definition of the gradient vector from the two points of view.



**Figure 25-15 (repeated)** Computer plot of the field lines and equipotentials for a charge distribution consisting of a positive charge + 3 and a negative charge - 1. These lines were then used to construct the plywood model.

## SLOPE IN TWO DIMENSIONS

Imagine that you are planning a trip in a desert with hills and valleys. One possibility is to follow a path that heads due east through the desert. If you draw the path on a contour map, and note where the path crosses different contours, you can create a plot of the height (h) of the path as a function of the distance (x) of the path. The result might look like a plot of h(x) shown in Figure (1). This should at least represent a smoothed version of the terrain you will encounter.

Your curve h(x) tells you roughly how steep the path should be at any point  $x_0$ . Mathematically, you can define the steepness as the slope of the tangent line at the point  $x_0$ , which is equal to the first derivative of h(x).

slope at 
$$x_0 = h'(x_0)$$
  
=  $\frac{dh(x)}{dx}\Big|_{x = x_0}$   
=  $\tan \theta$  (1)

As long as you stay on the path, the slope at any point is uniquely determined by Equation (1).



#### Figure 1

Imagine that you are walking due east (x direction) in the desert. We will call h(x) the height of your path. At some point  $x_0$ , the slope of your path is dh(x)/dxevaluated at  $x_0$ , which is the tangent of the angle  $\theta$ .

However the interesting part about going out in the desert is that you do not have to follow any particular path. If you do not want to climb very much, you can walk along a contour line. If you are anxious to get to the top of a hill and want the steepest climb possible, you walk at right angles to a contour line, along what we have called a *field line*, or what ski instructors call the *fall line*. At any point you can choose a path whose slope ranges from zero along a contour line to the maximum along the field line. To define the slope at some point, you have to state the direction you are traveling.

To handle this new feature mathematically, we first introduce a coordinate system (x,y), where the x direction, for example, could be east-west and the y direction north-south. The terrain is then described by a function h(x,y) giving the height of the land at any point (x,y).

To describe the slope of a one dimensional curve h(x) at some point  $x_0$ , we drew a *tangent line* at  $x_0$  as shown in Figure (1). To describe slopes for a two dimensional function h(x,y), at some point  $(x_0, y_0)$  we look at the *tangent plane* at that point. This assumes that the function h(x,y) is smooth enough that, as we get closer and closer to the point  $(x_0, y_0)$  the landscape looks smoother and smoother. It assumes that when we get very close, the landscape looks flat and we are looking at the tangent plane.

Not all functions h(x,y) are necessarily that smooth. Curves describing real landscapes, like the shape of a coastline, look just as rough no matter how close we look. Such curves are described by what is called *fractal geometry*. What we will be discussing are curves, or surfaces that become smooth when we look close enough. A sufficient mathematical criteria for such smoothness is that all derivatives with respect to any variable are finite.

If the terrain h(x,y) is smooth enough to have a unique tangent plane at every point, then our discussion of the nature of slopes on a curved surface can begin with a study of how slopes behave in a tangent plane. What we learn from the study of one tangent plane can then be applied to all tangent planes in the terrain. To visualize a tangent plane at some point  $(x_0, y_0)$ , start by imagining that the point is on the surface of a table, and construct a coordinate axis (x,y,z) whose origin is at  $(x_0, y_0)$  as shown in Figure (2). The xy plane is the table surface and the z axis points straight up. Let us assume that the x axis faces east and the y axis north.

To represent a tangent plane, take a thin flat object like a piece of cardboard, and place it on the table surface, tilted at an angle  $\theta$  as shown in Figure (3). Orient the cardboard so that the line of contact with the table is the x axis.

It is easy to see that in our flat tilted surface, all lines parallel to the x axis are contour lines, and that all lines parallel to the y axis headed north are field lines with a maximum slope. It is also clear that the field lines are perpendicular to the contour lines.

These features carry over to a smooth curved surface h(x,y). At any point  $(x_0, y_0)$  construct a tangent

plane. Unless this tangent plane happens to be horizontal, there will be a unique horizontal line in the plane that passes through the point  $(x_0, y_0)$ . This horizontal line corresponds to the x axis in Figure (3). In a region very close to the point  $(x_0, y_0)$  this horizontal line will coincide with the contour line of h(x,y) that passes through that point.

Perpendicular to the x axis in the tangent plane will be a line of maximum slope heading in the y direction of Figure (3). The field line of our curved surface h(x,y) that passes through the point  $(x_0, y_0)$ will be y oriented for a small region around  $(x_0, y_0)$ . As a result, in this small region the contour lines and the field lines of the curved surface have the same properties as the contour and field lines in the tangent plane. In particular, even for curved surfaces, contour lines and field lines will always be perpendicular to each other where the contour lines are in the direction of zero slope and the field lines in the direction of maximum slope.



Our coordinate system.

The tangent plane. All lines in the tangent plane that are parallel to the x axis are lines of equal height, or contour lines. Lines in the perpendicular v direction are lines of maximum slope, or field lines.

# THE GRADIENT

When you have a mathematical function h(x,y) that describes a surface, the slope of that surface in some direction is given by the partial derivative in that direction. Explicitly the slope in the x direction at the point  $(x_0, y_0)$  is given by

$$\left. \begin{array}{c} \text{slope in } x \\ \text{direction at} \\ (x_0, y_0) \end{array} \right\} = \left. \frac{\partial h(x, y)}{\partial x} \right|_{x = x_0, y = y_0}$$
(2a)

and the slope in the y direction is

slope in y  
direction at  
$$(x_0, y_0)$$
 =  $\frac{\partial h(x, y)}{\partial y}\Big|_{x = x_0, y = y_0}$  (2b)

What we will do now is to define a quantity we will call the *gradient*, and represent it by the symbol  $\vec{\nabla}h(x,y)$ . Explicitly we define  $\vec{\nabla}h(x,y)$  by the equation

$$\vec{\nabla}h(x,y) = \hat{x} \frac{\partial h(x,y)}{\partial x} + \hat{y} \frac{\partial h(x,y)}{\partial y}$$
 (3)

where  $\hat{x}$  and  $\hat{y}$  are unit vectors in the x and y directions respectively.

The gradient  $\vec{\nabla}h(x,y)$  looks like a vector with x and y components equal to the slope of h(x,y) in the x and y directions. However a vector is more than a quantity with some components. We saw in Chapter 2 of the Physics text that a vector has a basic physical significance that does not depend upon the coordinate system used to define the vector. What we need to do for our gradient is to find the basic significance of the quantity  $\vec{\nabla}h(x,y)$  and then show that the physical picture does not change when the gradient is evaluated in a different coordinate system. To see the physical significance of the gradient, we will evaluate  $\nabla h(x,y)$  at some point  $(x_0, y_0)$ , using a coordinate system where the x axis is parallel to the contour line passing through that point. That is the same coordinate system we used in our discussion of the tangent plane in Figure (3). Since the x axis lies along a contour line at the point of interest, there is no change in height as we move a short distance in the x direction, and thus the partial derivative in the x direction is zero.

$$\frac{\partial h(x,y)}{\partial x}\Big|_{\substack{x = x_0 \\ y = y_0}} = 0 \qquad \begin{array}{c} \text{for an x axis} \\ \text{lying along} \\ \text{a contour line} \end{array}$$
(4)

What remains of the gradient is

$$\vec{\nabla}h(x,y)\Big|_{\substack{x = x_0 \\ y = y_0}} = \hat{y}\frac{\partial h(x,y)}{\partial y}\Big|_{\substack{x = x_0 \\ y = y_0}} \quad \begin{array}{c} \text{for an x axis} \\ \text{lying along} \\ \text{a contour line} \end{array}$$
(5)

For this coordinate system, the gradient is purely y oriented, which is the direction of the field line through  $(x_0, y_0)$ . Also the magnitude of the gradient is equal to the magnitude of the steepest slope at  $(x_0, y_0)$ . As a result, physical significance of the gradient, at least in this special coordinate system, is that it *describes both the direction and magnitude of the steepest slope*.

Thus the gradient has both a magnitude and a direction like the displacement vectors we discussed in Chapter 2 of the Physics text. *If the components of the gradient change (transform) in the same way as the components of a displacement vector, then the magnitude and direction will be preserved when we go to a new (rotated) coordinate system.* The components will look different, but the magnitude and direction will be unchanged. To see whether the components of the gradient transform (change) like the components of a displacement vector, let us first review what happens to a purely y oriented displacement vector  $\vec{B}$  when we go to a new coordinate system (x',y') that is rotated by an angle  $\phi$  about the z axis as shown in Figure (4). You can easily see that in the x',y' coordinate system, the components of  $\vec{B}$  are

$$B_{x'} = B \sin \phi$$
  

$$B_{y'} = B \cos \phi$$
(6)

#### Exercise 1

(a) Show that for a purely x oriented vector  $\vec{A}$  the components of  $\vec{A}$  in the rotated (x', y') coordinate system are

$$A_{\chi'} = A \cos \phi$$

$$A_{\chi'} = -A \sin \phi$$
(7)

(b) Now show that if you start with a vector

$$\vec{C} = \hat{x}A + \hat{y}B \equiv \hat{x}C_x + \hat{y}C_y$$

which has components  $C_x = A$  in the x direction and  $C_y = B$  in the y direction, then in the rotated coordinate system, the components of  $\vec{C}$  are

$$C_{x'} = +C_x \cos \phi + C_y \sin \phi$$

$$C_{y'} = -C_x \sin \phi + C_y \cos \phi$$
(8)

(Equations (8) are the general formula for the transformation of the x and y components of a vector when we rotate the coordinate system by an angle  $\phi$  about the z axis.)



Figure 4

When we rotate the coordinate system about the z axis, the y directed vector  $\vec{B}$  gets components in both the x' and y' directions.

When we go from the coordinate system (x,y) to the rotated coordinate system (x',y'), the gradient

$$\vec{\nabla}h(x,y) = \hat{x}\frac{\partial h(x,y)}{\partial x} + \hat{y}\frac{\partial h(x,y)}{\partial y}$$
 (3) repeat

becomes

$$\vec{\nabla}h(\mathbf{x}',\mathbf{y}') = \hat{\mathbf{x}}'\frac{\partial h(\mathbf{x}',\mathbf{y}')}{\partial \mathbf{x}'} + \hat{\mathbf{y}}'\frac{\partial h(\mathbf{x}',\mathbf{y}')}{\partial \mathbf{y}'}$$
(9)

To calculate the new components  $\partial h(x',y')/\partial x'$  and  $\partial h(x',y')/\partial y'$  at some arbitrary point (x,y) we will use our familiar tangent plane of Figure (3) reproduced here as Figure (5). We have also drawn in the rotated coordinate system (x',y') seen in the top view of Figure (5). The coordinate axes x, y and x',y' all lie in the table top surface, what we can call the z = 0 plane.

The partial derivative, for example  $\partial h(x,y)/\partial y$ , represents the rate of change of the height h as we go out along the y axis. For the rotated coordinate system, the partial derivative  $\partial h(x',y')/\partial x'$  represents the rate of change of the height h as we go out along the x' axis. We will use these ideas to calculate the height  $\Delta h$  of the point A shown in Figure (5), a point that is a distance  $\Delta x'$  down the x' axis.

z directed straight up



#### Figure 5

Our tangent plane of Figure (3) showing the rotated coordinate system x',y', and the point A, a distance  $\Delta x'$  down the x' axis.

There are two distinct ways to get to the point A. One is to go down the x' axis directly, a distance  $\Delta x'$ . For this route we get as the formula for  $\Delta h$ 

$$\Delta h = \begin{pmatrix} \text{slope in} \\ \text{the } x' \\ \text{direction} \end{pmatrix} \times \begin{pmatrix} \text{distance we} \\ \text{go in the} \\ x' \text{ direction} \end{pmatrix}$$
$$\Delta h = \frac{\partial h(x',y')}{\partial x'} \times \Delta x'$$
(10)

The other way to get to point A is to go down the x axis a distance  $\Delta x$ , gaining a height  $\Delta h_x$  given by

$$\Delta h_{x} = \frac{\partial h(x,y)}{\partial x} \Delta x \tag{11}$$

and then go out a distance  $\Delta y$  in the old y direction, giving us an additional height  $\Delta h_v$  given by

$$\Delta h_{y} = \frac{\partial h(x,y)}{\partial y} \Delta y$$
 (12)

The height  $\Delta h$  at point A will be the sum of these two heights

$$\Delta h = \Delta h_{x} + \Delta h_{y}$$
$$= \frac{\partial h(x,y)}{\partial x} \Delta x + \frac{\partial h(x,y)}{\partial y} \Delta y$$
(13)



**Figure 5 repeated** 

Our tangent plane of Figure (3) showing the rotated coordinate system x', y', and the point A, a distance  $\Delta x$  down the x' axis.

(In our drawing of Figure (5), we have shown the x axis as being horizontal, so that the slope  $\partial h(x,y)/\partial x$  would be zero. This makes the drawing easier to interpret, but we do not need to assume the x slope is zero for the current discussion.)

The final step in calculating the height  $\Delta h$  of point A from the second route is to relate  $\Delta x$  and  $\Delta y$  to the distance  $\Delta x'$  traveled along the x' axis. From the top view of Figure (5) it is clear that

$$\Delta x = \Delta x' \cos \phi$$
  
$$\Delta y = \Delta x' \sin \phi \qquad (14)$$

Using these values in Equation (13) give us

$$\Delta h = \frac{\partial h(x,y)}{\partial x} \Delta x' \cos \phi + \frac{\partial h(x,y)}{\partial y} \Delta x' \sin \phi$$
(15)

We can now equate our two formulas, Equation (10) and Equation (15) for the height  $\Delta h$  at point A. The factors of  $\Delta x'$  cancel and we are left with

$$\frac{\partial h(x',y')}{\partial x'} = \frac{\partial h(x,y)}{\partial x}(\cos \phi) + \frac{\partial h(x,y)}{\partial y}(\sin \phi)$$
(16)

Comparing Equation (15) with Equation (8) for the transformation of the x component of the displacement vector  $\vec{C}$ 

$$C_{x'} = C_x \cos \phi + C_y \sin \phi$$
 (8a) repeated

we see that the x component of the gradient transforms (changes) in the same way as a displacement vector when we rotate the coordinate system by an angle  $\phi$ .

#### Exercise 2

Using similar arguments, show that the y' slope  $\partial h(x',y')/\partial y'$  is given by

$$\frac{\partial h(x',y')}{\partial y'} = \frac{\partial h(x,y)}{\partial x} \left(-\sin\phi\right) + \frac{\partial h(x,y)}{\partial y} \left(\cos\phi\right) (17)$$

which is the same as the transformation of the y component of a displacement vector.

#### **Gradient as a Vector Field**

What is the significance of our demonstration that the quantity  $\vec{\nabla}h(x,y)$ , defined by

$$\vec{\nabla}h(x,y) = \hat{x}\frac{\partial h(x,y)}{\partial x} + \hat{y}\frac{\partial h(x,y)}{\partial y}$$
 (3) Repeat

transforms like a vector at each point (x,y) in space? As we pointed out in Chapter 29 of the Physics text, a vector field, which is a vector at every point in space, is uniquely determined if we have general formulas for the surface integral and the line integral of the field. There were four Maxwell's equations because we needed formulas for the surface and the line integrals of both the electric and magnetic fields.

In the Physics text and the first part of this chapter, we knew that the electric field was a vector field because of its definition as the force vector acting on a unit test charge. The knowledge that forces transform as vectors was sufficient to tell us that any correct formula for  $\vec{E}$  gave us a vector field. In this section with the definition of Equation (3), the gradient is given a geometrical definition, which at first sight might or might not make  $\vec{\nabla}h(x,y)$  behave as a vector field. The demonstration that  $\vec{\nabla}h(x,y)$ transforms as a vector means that concepts like line and surface integrals can be applied to any gradient fields. As we saw in the first part of this chapter, the extension of Equation (3) to the gradient of a three dimensional function is

$$\vec{\nabla}h(x,y,z) = \hat{x}\nabla_{x}h + \hat{y}\nabla_{y}h + \hat{z}\nabla_{z}h \qquad (18)$$

where  $\nabla_x$ ,  $\nabla_y$  and  $\nabla_z$  are the partial derivatives  $\partial/\partial x$ ,  $\partial/\partial y$ , and  $\partial/\partial z$ . Equation (18) here is equivalent to Equation (16) in the first part of the chapter relating  $\vec{E}$  to  $\vec{\nabla}V(x,y,z)$ .

This completes our discussion of the gradient vector  $\vec{\nabla}h(x,y,z)$  from a geometrical point of view. If you have not done so already, now is the time to look at applications of the gradient vector to electric field problems, starting with the discussion of the gradient vector just before Equation (16) of the first part of the chapter.

# View 3 **Pressure Force as a Gradient**

We end the chapter with View 3, an application to fluids, where we see that the pressure force  $\vec{f}_p$  acting on the fluid particles is the gradient of the pressure field p. This represents a straightforward example of obtaining a vector field  $\vec{f}_p$  from a scalar field p.

# PRESSURE FORCE AS A GRADIENT

In the Physics text, there were two main places where we dealt with the concept of pressure. The first was in Chapters 17 and 18 on the ideal gas law, and the second was in Chapter 23 during our discussion of Bernoulli's equation. In both cases we mentioned that pressure had the dimensions of a force per unit area, but was itself a scalar field p(x,y, z) that did not point anywhere. We pointed out that the pressure force acting on an area  $\Delta A$  was directed perpendicular to the area and had a magnitude

$$\Delta F = p \Delta A \tag{1}$$

We will now use the concept of a gradient to show that the pressure force per unit volume  $\vec{f}_p$ , acting on the fluid particles, is equal to minus the gradient of the pressure p(x,y,z)

$$\vec{f}_{p} = -\vec{\nabla}p(x,y,z) \tag{2}$$

This is analogous to the electric field being equal to minus the gradient of the electric voltage

$$\vec{E} = -\vec{\nabla}V(x,y,z) \tag{3-19}$$

which we saw back in Equation (3-19) of this chapter.

To calculate the pressure force, we start with a small volume  $\Delta V = \Delta x \Delta y \Delta z$  shown in Figure (1). This volume element has a left face located at z and a right face at  $z + \Delta z$ . The center of the faces are located at (x, y) where the pressures are p(x, y, z) and  $p(x,y,z + \Delta z)$  respectively.

The pressure force  $\Delta \vec{F}_1$  exerted on the left face of  $\Delta V$  is equal to the force per unit area  $p_1(x,y,z)$  times the area  $\Delta A_1 = \Delta x \Delta y$  of that face. The pressure force is directed into the volume, toward the right in the  $\hat{z}$  direction, as shown

$$\Delta \dot{F}_1 = \hat{z} p(x, y, z) \Delta x \Delta y \tag{3}$$

On the right side, the force is directed back into  $\Delta V$ , in the -z direction, and has a value

$$\Delta \dot{F}_2 = -\hat{z}p(x,y,z + \Delta z)\Delta x \Delta y \tag{4}$$

The net force on these two sides is

$$\begin{split} \left(\Delta \vec{F}_{1} + \Delta \vec{F}_{2}\right) \\ &= -\hat{z} \Big[ p(x,y,z + \Delta z) - p(x,y,z) \Big] \Delta x \Delta y \\ &= -\hat{z} \Big[ \frac{p(x,y,z + \Delta z) - p(x,y,z)}{\Delta z} \Big] \Delta x \Delta y \Delta z \end{split}$$
(5)

You can immediately see that when we take the limit that  $\Delta V$  is an infinitesimal volume and  $\Delta z$  goes to zero, the quantity in the square brackets in Equation (5) becomes the partial derivative of p(x,y,z) with respect to z.



Figure 1 The volume element  $\Delta x \Delta y \Delta z$ .

$$\lim_{\Delta z \to 0} \left[ \frac{p(x,y,z + \Delta z) - p(x,y,z)}{\Delta z} \right] = \frac{\partial p(x,y,z)}{\partial z}$$
(6)

Thus Equation (5) can be written in the somewhat mixed form

$$\left(\Delta \vec{F}_1 + \Delta \vec{F}_2\right) = -\hat{z} \frac{\partial p(x, y, z)}{\partial z} \Delta x \Delta y \Delta z \qquad (7)$$

where we will shortly think in terms of the limit that  $\Delta V = \Delta x \Delta y \Delta z$  goes to zero.

Before we do, let us add in the pressure forces  $F_3$ and  $\vec{F}_4$  acting on the bottom and top faces respectively, and  $\vec{F}_5$  and  $\vec{F}_6$  acting on the back and front faces to get the total pressure force  $\Delta \vec{F}_p$  acting on  $\Delta V$ . Following the same steps used to derive Equation (7), we get

$$\begin{split} \Delta \vec{F}_{p} &= \left( \Delta \vec{F}_{1} + \Delta \vec{F}_{2} + \Delta \vec{F}_{3} + \Delta \vec{F}_{4} + \Delta \vec{F}_{5} + \Delta \vec{F}_{6} \right) \\ &= - \left[ \hat{z} \frac{\partial p(x, y, z)}{\partial z} + \hat{y} \frac{\partial p(x, y, z)}{\partial y} + \hat{x} \frac{\partial p(x, y, z)}{\partial x} \right] \Delta x \Delta y \Delta z \end{split}$$

$$\end{split}$$

$$(8)$$

The quantity in the square brackets in Equation (8) is the gradient  $\overline{\nabla}p$  of the pressure field. Thus we have, after dividing both sides by  $\Delta V = \Delta x \Delta y \Delta z$ 

$$\frac{\Delta \vec{F}_{p}}{\Delta V} = -\vec{\nabla} p(x, y, z)$$
(9)

We recognize the left side of Equation (9) as the total pressure force acting on  $\Delta V$  divided by the volume  $\Delta V$ . It is therefore the *pressure force per unit volume*  $\vec{f}_{p}(x,y,z)$  acting in that region of the fluid, and we get our advertized result

$$\vec{f}_{p}(x,y,z) = -\vec{\nabla}p(x,y,z)$$
  
*pressure force per unit volume*

(2) repeated

olume

With Equation (2) we have a powerful way of calculating pressure forces, since we can evaluate the gradient in any of the coordinate systems we have been discussing, such as cylindrical or spherical polar coordinates.

# Calculus 2000-Chapter 4 The Operator $\nabla^2 \equiv \vec{\nabla} \cdot \vec{\nabla}$ (The Laplacian)

In our earliest discussion of vectors in Chapter 2 of the Physics text, we were introduced to the vector dot product

$$\vec{A} \cdot \vec{B} \equiv A_x B_x + A_y B_y + A_z B_z \tag{1}$$

as having the special property of being a scalar quantity. That is, the quantity  $\vec{A} \cdot \vec{B}$  had the same value no matter what coordinate system we used to evaluate it. Having just seen that the gradient operator  $\vec{\nabla}$  operating on a scalar field h(x,y,z)produces a vector field, one might wonder what we get when we take the dot product of two gradient operations acting on a scalar field. The answer is that we get another scalar field.

The standard name for this dot product of two gradient operators is **del squared**, written as

$$\nabla^2 \equiv \vec{V} \cdot \vec{V} \tag{2}$$

It is often called the **Laplacian** operator after the French mathematician Laplace. This operator is essentially an extension to three dimensions of the

second derivative we encountered in Calculus Chapter 2, during our discussion of the one dimensional wave equation. Thus we should expect  $\nabla^2$  to appear when we begin to discuss three dimensional wave equations in the next few chapters.

## **Fluid theory**

Another area of physics where the operator  $\nabla^2$ plays a prominent role is in fluid dynamics. For common fluids like water and air, the viscous force acting on the fluid particles turns out to be proportional to the Laplacian of the velocity field, namely  $\nabla^2 \vec{v}$ . We will derive that result starting from an assumption that Issac Newton made about the nature of viscous forces.

As an application of the theory of viscous forces, we will look at the steady flow of a viscous fluid in a pipe. This example provides a way to measure the so called **coefficient of viscosity** that appears in Newton's theory. It also provides an example of the use of the operator  $\nabla^2$  acting on a vector field.

#### Schrödinger's Equation

One of the glaring omissions in the Physics text resulted from our inability to calculate the electron wave patterns in the hydrogen atom. All we were able to do is show drawings of a few of the lowest energy wave patterns, describe the electron's energy and angular momentum in these wave patterns, and then state that these patterns came from a wave equation called Schrödinger's equation. We were neither able to write down or solve the equation itself.

To handle Schrödinger's equation as applied to the hydrogen atom, we needed two mathematical concepts we did not then have. One is the operator  $\nabla^2$  which we are introducing in this chapter, the other is the concept of a **complex variable** which we will introduce in the next chapter, Chapter 5. Once we develop these two mathematical tools, we will be ready to approach Schrödinger's equation in Chapter 6.

When we apply Schrödinger's equation to the hydrogen atom, we are dealing with a system that has spherical symmetry. As a result it is much easier to deal with the theory using a coordinate system that has the same symmetry. The problem is that the operator  $\nabla^2$ , which in Cartesian coordinates is a straightforward extension of the second derivative, becomes quite complex when we work in other coordinate systems like spherical polar coordinates. The reason for the complexity is that in any coordinate system except Cartesian coordinates, the unit vectors may change direction as we move from one point in space to another. This change in the direction of the unit vectors complicates the formulas for  $\nabla^2$ .

#### **The Formulary**

In the main part of this chapter we will simply state the formula, in spherical polar coordinates, for  $\nabla^2$  acting on a scalar field  $\psi$ . This is the formula we will use in Chapter 6 in our discussion of the hydrogen atom. In the appendix, however, we will derive the formula, showing you exactly how the changing unit vectors affect the results. We have placed this derivation in an appendix because it is the kind of derivation you probably want to observe only once in your life, to find out where the rather messy results come from.

When you are actually working problems involving quantities like  $\nabla^2$  in cylindrical or spherical coordinates, you do not want to derive the formulas yourself because the chances of your getting the right answer are too small. You are not likely to memorize them correctly either, unless you use a particular formula often. Instead, the best procedure is to look up the result in a table of formulas, sometimes called a **formulary**. We provide a formulary at the end of this text, one adapted from a formulary developed by David Book of the Naval Research Laboratory.

In our discussion of viscous forces in this chapter, we use the formulary to find the formula in cylindrical coordinates for  $\nabla^2$  acting on the vector field  $\vec{v}$ .

# $\nabla^2$ IN CARTESIAN COORDINATES

We will first take a careful look at  $\nabla^2 = \vec{\nabla} \cdot \vec{\nabla}$  in Cartesian coordinates before we approach the spherical case. Using the unit vector notation for  $\vec{\nabla}$  we have

$$\vec{\nabla} = \hat{x}\frac{\partial}{\partial x} + \hat{y}\frac{\partial}{\partial y} + \hat{z}\frac{\partial}{\partial z}$$
(3)

where  $\hat{x}$ ,  $\hat{y}$  and  $\hat{z}$  are unit vectors pointing in the x, y, and z directions respectively. The dot product  $\vec{\nabla} \cdot \vec{\nabla}$  acting on some function f(x,y,z) should be given by

$$\begin{aligned} \vec{\nabla} \cdot \vec{\nabla} f(\mathbf{x}, \mathbf{y}, \mathbf{z}) \\ &= \left( \hat{\mathbf{x}} \frac{\partial}{\partial \mathbf{x}} + \hat{\mathbf{y}} \frac{\partial}{\partial \mathbf{y}} + \hat{\mathbf{z}} \frac{\partial}{\partial \mathbf{z}} \right) \cdot \left( \hat{\mathbf{x}} \frac{\partial f}{\partial \mathbf{x}} + \hat{\mathbf{y}} \frac{\partial f}{\partial \mathbf{y}} + \hat{\mathbf{z}} \frac{\partial f}{\partial \mathbf{z}} \right) \\ &= \left( \hat{\mathbf{x}} \frac{\partial}{\partial \mathbf{x}} \right) \cdot \left( \hat{\mathbf{x}} \frac{\partial f}{\partial \mathbf{x}} \right) + \left( \hat{\mathbf{x}} \frac{\partial}{\partial \mathbf{x}} \right) \cdot \left( \hat{\mathbf{y}} \frac{\partial f}{\partial \mathbf{y}} \right) + \cdots \\ &+ \cdots + \left( \hat{\mathbf{z}} \frac{\partial}{\partial \mathbf{z}} \right) \cdot \left( \hat{\mathbf{z}} \frac{\partial f}{\partial \mathbf{z}} \right) \end{aligned}$$
(4)

Being very careful with our differentiation, we have, for example,

$$\left(\hat{\mathbf{x}}\frac{\partial}{\partial \mathbf{x}}\right) \cdot \left(\hat{\mathbf{x}}\frac{\partial \mathbf{f}}{\partial \mathbf{x}}\right) = \hat{\mathbf{x}} \cdot \left[\frac{\partial \hat{\mathbf{x}}}{\partial \mathbf{x}}\frac{\partial \mathbf{f}}{\partial \mathbf{x}} + \hat{\mathbf{x}}\frac{\partial^2 \mathbf{f}}{\partial \mathbf{x}^2}\right]$$
(5)

We have been overly careful because the unit vectors  $\hat{x}$ ,  $\hat{y}$  and  $\hat{z}$  are constant in both magnitude and direction, thus

$$\frac{\partial \hat{\mathbf{x}}}{\partial \mathbf{x}} = 0 \tag{6}$$

and we are left with

$$\left(\hat{\mathbf{x}}\frac{\partial}{\partial \mathbf{x}}\right) \cdot \left(\hat{\mathbf{x}}\frac{\partial \mathbf{f}}{\partial \mathbf{x}}\right) = \hat{\mathbf{x}} \cdot \hat{\mathbf{x}}\frac{\partial^2 \mathbf{f}}{\partial \mathbf{x}^2} = \frac{\partial^2 \mathbf{f}}{\partial \mathbf{x}^2} \tag{7}$$

Similarly

$$\left( \hat{\mathbf{x}} \frac{\partial}{\partial \mathbf{x}} \right) \cdot \left( \hat{\mathbf{y}} \frac{\partial \mathbf{f}}{\partial \mathbf{y}} \right) = \hat{\mathbf{x}} \cdot \left( \frac{\partial \hat{\mathbf{y}}}{\partial \mathbf{x}} \frac{\partial \mathbf{f}}{\partial \mathbf{y}} + \hat{\mathbf{y}} \frac{\partial^2 \mathbf{f}}{\partial \mathbf{x} \partial \mathbf{y}} \right)$$
$$= \hat{\mathbf{x}} \cdot \hat{\mathbf{y}} \frac{\partial^2 \mathbf{f}}{\partial \mathbf{x} \partial \mathbf{y}} = 0$$
(8)

because  $\partial \hat{y} / \partial x = 0$  and  $\hat{x} \cdot \hat{y} = 0$ .

As a result, all we are left with, when we evaluate  $\nabla^2 f$  in Cartesian coordinates is

$$\nabla^{2} \mathbf{f}(\mathbf{x},\mathbf{y},\mathbf{z}) = \frac{\partial^{2} \mathbf{f}}{\partial \mathbf{x}^{2}} + \frac{\partial^{2} \mathbf{f}}{\partial \mathbf{y}^{2}} + \frac{\partial^{2} \mathbf{f}}{\partial \mathbf{z}^{2}}$$
(9)

which is an obvious extension to three dimensions of the second derivative  $\partial^2 f/\partial x^2$  that appeared in our one dimensional wave equation in Chapter 2 of the Calculus text.

# $abla^2$ in Spherical Polar Coordinates

As we mentioned, the results are not so simple when we are working in other coordinate systems. In spherical polar coordinates, when  $\nabla^2$  is acting on a scalar function, we get the following result which is derived in the appendix to this chapter.

$$\nabla^{2} \mathbf{f} = \frac{1}{r} \frac{\partial^{2}}{\partial r^{2}} (\mathbf{r} \mathbf{f}) + \frac{1}{r^{2}} \left[ \frac{1}{\sin \theta} \frac{\partial}{\partial \theta} \left( \sin \theta \frac{\partial \mathbf{f}}{\partial \theta} \right) + \frac{1}{\sin^{2} \theta} \frac{\partial^{2} \mathbf{f}}{\partial \phi^{2}} \right]$$
(10)

(10)

where r,  $\theta$ , and  $\phi$  are the polar coordinates shown in Figure (1). Much of this complexity comes from the fact that the unit vectors are not constant, and have to be differentiated. You will see how this works by going to the appendix.

(We should note that, in non Cartesian coordinates,  $\nabla^2$  acting on a vector, e.g.  $\nabla^2 \vec{E}$ , has an even more complex formula, which is given in the formulary at the end of the text.)



Spherical polar coordinates.

# **NEWTONIAN FLUIDS**

We now move on to our example of the use of the Laplacian operator to describe viscosity in a Newtonian fluid.

Newton proposed that viscous effects in a fluid resulted from the shearing motion of one layer of fluid over another. This shearing force can be introduced as follows.

Suppose we have a simple flow where all the fluid is moving in the x direction, and the velocity is increasing in the y direction as shown in Figure (2).

To analyze the forces involved, consider a horizontal plane indicated by the dashed line labeled by A----B. The fluid above the plane, which is travelling faster, drags the fluid below forward. The fluid below, which is going slower, drags the upper fluid back. Let  $\vec{\tau}_+$  be the force per unit area exerted by the upper fluid on the lower fluid, and  $\vec{\tau}_-$ , the force exerted by the lower fluid on the upper. In Figure (2) we have drawn the forces  $\vec{\tau}_+$  and  $\vec{\tau}_-$  inside the fluids upon which they act.



**Figure 2** Diagram of a simple flow where the velocity field  $\vec{v}$  is x directed and increasing in the y direction.

This combination of oppositely directed forces on opposite sides of the plane is called a *stress*, in this case a stress generated by the action of viscosity. For a so called *Newtonian fluid*, the stress  $\tau$  is assumed to be directly proportional to the rate at which the velocity field is changing as we move up, which for our x directed flow is

$$\tau = \mu \frac{\partial v_x(y)}{\partial y}$$
(11)

The quantity  $\mu$  is called the *coefficient of viscosity* 

$$\mu$$
 = coefficient of viscosity (12)

For a Newtonian fluid,  $\mu$  is assumed to be a constant throughout the fluid. In many situations, both water and air behave as Newtonian fluids.

#### **VISCOUS FORCE ON A FLUID ELEMENT**

Suppose again that we have a simple x directed velocity field whose velocity profile is shown in Figure (3). Now consider a small volume element with sides  $\Delta x$ ,  $\Delta y$  and  $\Delta z$ , the bottom of which is located at (y) and the top at (y +  $\Delta y$ ) is shown. The fluid below the plane A----B at y is dragging the fluid above, back with a force per unit area  $\tau_{-}(y)$ 

$$\tau_{-}(y) = -\mu \frac{\partial v_{x}(y)}{\partial y} \quad \begin{array}{c} \text{force per unit} \\ \text{area at the bottom} \\ \text{of volume element} \end{array} \tag{13}$$

The total force at the bottom is the force per unit area  $\tau_{-}(y)$  times the area  $\Delta x \Delta z$  upon which it is acting

$$\Delta F_{-}(y) = \tau_{-}(y)\Delta x\Delta z$$

$$= -\mu \frac{\partial v_{x}(y)}{\partial y}\Delta x\Delta z$$
(14)

Up at the top of the volume element, the faster fluid above the C----D plane at  $(y + \Delta y)$ , is pulling forward the slower fluid below with a total force

With Equations (14) and (15) we see that the total viscous force on the fluid in our volume element can be written

$$\Delta F_{x} = \Delta F_{-}(y) + \Delta F_{+}(y + \Delta y)$$
$$= \left[ -\mu \frac{\partial v_{x}(y)}{\partial y} + \mu \frac{\partial v_{x}(y + \Delta y)}{\partial y} \right] \Delta x \Delta z$$
(16)

Multiplying the right side by  $\Delta y / \Delta y$  gives

$$\Delta F_{x} = \mu \left[ \frac{\frac{\partial v_{x}(y + \Delta y)}{\partial y} - \frac{\partial v_{x}(y)}{\partial y}}{\Delta y} \right] \Delta x \Delta y \Delta z \ (17)$$

The quantity in the square brackets should be recognized as the second derivative of  $v_x(y)$  with respect to y. Dividing through both sides by the volume  $\Delta x \Delta y \Delta z$  gives us the viscous force per unit volume

$$\frac{\Delta F_{x}}{\Delta x \Delta y \Delta z} = f_{vx} \qquad \begin{array}{l} viscous \ force \ per \ unit \\ volume \ acting \ on \\ the \ fluid \ element \end{array}$$

$$f_{vx} = \mu \frac{\partial^{2} v_{x}(y)}{\partial y^{2}} \qquad (18)$$

This is the formula for the viscous force per unit volume acting on the fluid particles when we have a purely x directed flow of a Newtonian fluid whose speed varies only in the y direction. In the next section we generalize the result to three dimensional flows.

**Figure 3** *Calculating the viscous force on a fluid element.* 

# VISCOUS FORCE FOR THREE DIMENSIONAL FLOWS

At first sight, there seems to be a rather obvious extension of Equation (18) to three dimensional flows. In a chapter devoted to discussing the operator  $\nabla^2$ , we might expect that the generalization of our formula for the viscous force  $\vec{f}_v$  per unit volume should be

$$\vec{f}_{v} = \mu \nabla^{2} \vec{v} \tag{19}$$

To check that Equation (19) reduces to our result in Equation (18), when  $\vec{v}$  is the one dimensional flow  $v_x(y)$ , we have

$$f_{vx} = \mu \nabla^2 v_x(y) = \mu \left( \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2} \right) v_x(y)$$
$$= \mu \frac{\partial^2 v_x(y)}{\partial y^2}$$
(20)

Thus we get the desired result for one dimensional flows.

However, complications arise in three dimensional flows that we did not consider in our analysis of the simple one dimensional flow pattern. In three dimensions, fluids flow around corners and x directed flows can become y or z directed. The definition of viscous stress we gave in Equation (11) simply cannot handle changes in the direction of the flow.

An effective way to deal with viscous forces in three dimensional flows is to note that the resulting force  $\vec{f}_v$  per unit volume must be a vector field. That is,  $\vec{f}_v$  must transform like a vector field when we rotate the coordinate system. (See the discussion of the transformation of vector fields at the end of the geometrical discussion of the gradient in Chapter 3.)

We will also require that  $\vec{f}_v$  be made up of some combination of constants and second derivatives of the velocity field. These requirements on  $\vec{f}_v$  are essentially what we mean by a Newtonian fluid with constant coefficients. If the viscous forces are more complex, which they can be for something like a liquid crystal, then we say that the fluid is non Newtonian. What we want is the most general combination we can make out of constants, two derivatives  $\vec{\nabla}$ , and a velocity field  $\vec{v}$ . Basically we have three vectors  $\vec{\nabla}$ ,  $\vec{\nabla}$ ,  $\vec{v}$ , and we must multiply them together to get a single vector. To do this, we have to take the dot product of two of them. The possibilities are  $(\vec{\nabla} \cdot \vec{\nabla}) \vec{v}$  and  $\vec{\nabla} (\vec{\nabla} \cdot \vec{v})$ .\* As a result, our most general formula for a Newtonian fluid with constant coefficients is

$$\vec{f}_{\nu} = \mu_1 (\vec{\nabla} \cdot \vec{\nabla}) \vec{v} + \mu_2 \vec{\nabla} (\vec{\nabla} \cdot \vec{v})$$
(21)

where  $\mu_1$  and  $\mu_2$  are constants. There is no other combination of constants and second derivatives of the velocity field that transforms as a vector when we rotate the coordinate system.

If we are dealing with a constant density fluid,  $\vec{\nabla} \cdot \vec{v} = 0$  and we are left with

$$\vec{f}_{\nu} = \mu_1 (\vec{\nabla} \cdot \vec{\nabla}) \vec{v} = \mu \nabla^2 \vec{v}$$
(19a)

which is the result we guessed back in Equation (19), with  $\mu_1 = \mu$ .

Equation (21) suggests that it is possible to have a second kind of viscosity when the fluid is compressible and  $\vec{\nabla} \cdot \vec{v}$  is not zero. This has in fact been observed, and  $\mu_2$  is sometimes called the *second* viscosity coefficient. (Some texts use a second viscosity coefficient defined as  $\lambda = \mu_2 - \mu$ .) In this text we will only deal with incompressible fluids where there is no second viscosity, and  $\vec{f}_v$  is simply given by the Laplacian operator  $\nabla^2$  acting on  $\vec{v}$ , namely  $\vec{f}_v = \mu \nabla^2 \vec{v}$ .

\* (You might also consider vector cross products involving  $\vec{\nabla}$ ,  $\vec{\nabla}$ , and  $\vec{v}$ . The possibilities are  $\vec{\nabla} \times (\vec{\nabla} \cdot \vec{v})$ ,  $\vec{\nabla} \cdot (\vec{\nabla} \times \vec{v})$ , and  $\vec{\nabla} \times (\vec{\nabla} \times \vec{v})$ . At the beginning of Chapter 9, we find that the first two of these are identically zero, and the third turns out to be

$$\vec{\nabla} \times (\vec{\nabla} \times \vec{v}) = \vec{\nabla} (\vec{\nabla} \cdot \vec{v}) - (\vec{\nabla} \cdot \vec{\nabla}) \vec{v}$$

which involves only the two terms we got from dot products. Thus we get nothing new by considering cross products. )

#### Viscous Force in Cylindrical Coordinates

Now that we have the formula for the viscous force  $\vec{f}_v = \mu \nabla^2 \vec{v}$ , which applies to any fluid that we will consider in this text, we are free to use general formulas we have in the formulary for  $\nabla^2$  in various coordinate systems. We are about to study the flow of a viscous fluid in a pipe, a problem that obviously has cylindrical symmetry. Thus to analyze the viscous forces, we should work with  $\nabla^2 \vec{v}$  in cylindrical coordinates.

We mentioned earlier that  $\nabla^2$  acting on a vector field is more complex than  $\nabla^2$  acting on a scalar field in anything except Cartesian coordinates. Thus evaluating  $\nabla^2 \vec{v}$  in cylindrical coordinates will give us some practice in correctly using the formulary.

From the formulary we find the following formula for  $\nabla^2$  acting on a scalar field f and a vector field  $\vec{A}$ .

$$\nabla^{2} \mathbf{f} = \frac{1}{\mathbf{r}} \frac{\partial}{\partial \mathbf{r}} \left( \mathbf{r} \frac{\partial \mathbf{f}}{\partial \mathbf{r}} \right) + \frac{1}{\mathbf{r}^{2}} \frac{\partial^{2} \mathbf{f}}{\partial \theta^{2}} + \frac{\partial^{2} \mathbf{f}}{\partial z^{2}}$$
(22)

where the coordinates r,  $\theta$ , z are the unit vectors  $\hat{\mathbf{r}}$ ,  $\hat{\theta}$ ,  $\hat{z}$  shown in Figure (4).



**Figure 4** *Cylindrical coordinates*.

Looking farther down in the formulary we find for the components of  $\nabla^2 \vec{A}$ 

$$(\nabla^2 \vec{A})_r = \nabla^2 A_r - \frac{2}{r^2} \frac{\partial A_\theta}{\partial \theta} - \frac{A_r}{r^2}$$
(23a)

$$(\nabla^2 \vec{A})_{\theta} = \nabla^2 A_{\theta} + \frac{2}{r^2} \frac{\partial A_r}{\partial \theta} - \frac{A_{\theta}}{r^2}$$
(23b)

$$(\nabla^2 \vec{A})_z = \nabla^2 A_z \tag{23c}$$

where, for example,  $\nabla^2 A_z$  means apply Equation (22) to  $A_z$ 

$$\nabla^{2} \mathbf{A}_{z}(\mathbf{r}, \boldsymbol{\theta}, z) = \frac{1}{r} \frac{\partial}{\partial r} \left( r \frac{\partial \mathbf{A}_{z}}{\partial r} \right) + \frac{1}{r^{2}} \frac{\partial^{2} \mathbf{A}_{z}}{\partial \theta^{2}} + \frac{\partial^{2} \mathbf{A}_{z}}{\partial z^{2}}$$
(24)

All this looks like a terrible mess. But suppose we have a fluid flowing smoothly along a pipe as shown in Figure (5). Taking the  $\hat{z}$  direction down the pipe and r the distance out from the axis of the pipe, we can assume, for cylindrical symmetry, that  $\vec{v}(r,\theta,z)$  is purely  $\hat{z}$  directed and depends only on the radius r.

$$\vec{v}(r,\theta,z) = \hat{z}v_{z}(r) \tag{25}$$

Now let us work out  $\nabla^2 \vec{v}$  for this simple case using Equations (23) for  $\nabla^2$  in cylindrical coordinates. Because  $v_r$  and  $v_{\theta}$  are zero, we do not worry about Equations (23a) and (23b). From (23c) we have

$$\left(\nabla^2 \vec{\mathbf{v}}\right)_z = \nabla^2 \mathbf{v}_z \tag{26}$$

Thus for this case we do not have to worry about the extra stuff that comes in when we take  $\nabla^2$  of a vector.



**Figure 5** Velocity profile for the uniform flow in a pipe.

Next we note that  $v_z = v_z(r)$ , thus we can ignore the  $\partial v_z/\partial \theta$  and  $\partial v_z/\partial z$  terms in (13a) and we are left with

$$\left(\nabla^2 \vec{\mathbf{v}}\right)_z = \frac{1}{r} \frac{\partial}{\partial r} \left[ r \frac{\partial \mathbf{v}_z(r)}{\partial r} \right]$$
(27)

which is not such a difficult thing to work with after all.

To get a feeling for what the viscous force looks like for pipe flow, we look up in a fluids text what the so called *laminar* (i.e., non turbulent) velocity profile is in a pipe. The result they give is

$$v_{z}(\mathbf{r}) = \frac{V_{0}}{R^{2}} (R^{2} - r^{2}) \qquad \begin{array}{c} parabolic \\ velocity \\ profile \end{array} (28)$$

where R is the radius of the pipe,  $V_0$  the flow speed at the center, and r the radial distance from the axis. This is the parabolic profile shown in Figure (5). You can see that at the edge of the pipe, where r = R, the velocity goes to zero. At the center where r = 0,  $v_z = V_0$  is a maximum.

To calculate the viscous force per unit volume for this parabolic profile, we have

$$\vec{f}_{v} = \mu \nabla^{2} \vec{v} \tag{29}$$

$$(f_{v})_{z} = \mu (\nabla^{2} \vec{v})_{z} = \mu \frac{1}{r} \frac{\partial}{\partial r} \left( r \frac{\partial v_{z}(r)}{\partial r} \right)$$
 (30)

With Equation (28) written as

$$v_{z}(r) = -\frac{V_{0}}{R^{2}}r^{2} + V_{0}$$
(28a)

we easily get

$$\frac{\partial v_z(\mathbf{r})}{\partial \mathbf{r}} = -\frac{2V_0}{R^2} \mathbf{r}$$
$$\mathbf{r} \frac{\partial v_z(\mathbf{r})}{\partial \mathbf{r}} = -\frac{2V_0}{R^2} \mathbf{r}^2$$
(31)

Thus  $(\nabla^2 \vec{v})_z$  becomes  $(\nabla^2 \vec{v})_z = \frac{1}{r} \frac{\partial}{\partial r} \left( r \frac{\partial v_z(r)}{\partial r} \right)$   $= \frac{1}{r} \frac{\partial}{\partial r} \left( -\frac{2V_0}{R^2} r^2 \right)$   $= \frac{1}{r} \left( -\frac{2V_0}{R^2} \frac{\partial r^2}{\partial r} \right)$  (32)  $= \frac{1}{r} \left( -\frac{2V_0}{R^2} 2r \right)$ 

The r's cancel and we are left with

$$(\nabla^2 \vec{v})_z = -\frac{4V_0}{R^2}$$
(33)

The viscous force  $\vec{f}_v = \mu \nabla^2 \vec{v}$  becomes

$$f_{vz} = -\mu \left(\frac{4V_0}{R^2}\right)$$
(34)

We end up with the result that  $\vec{f}_v$  points in the  $-\hat{z}$  direction (it has only a negative z component) and is constant in magnitude throughout the pipe. This is a wonderfully simple result considering the staggering mess of terms we faced in Equation (23).

We will see that the physics of the parabolic laminar flow is that this uniform  $-\hat{z}$  oriented viscous force is balanced by a uniform  $+\hat{z}$  oriented pressure gradient down the tube. Thus there is no net force on each fluid element and the fluid moves down the pipe without acceleration, i.e., at constant velocity.

#### Measuring the Viscosity Coefficient

If we have an apparatus where we know the pressure gradient, we can use that to measure the viscosity coefficient  $\mu$  of the fluid. Such an apparatus is sketched in Figure (6), a sketch taken from the excellent fluid dynamics text by Tritton.

Since there is essentially no viscosity acting in the region between points (1) at the top of the fluid in the container, and point (2) near the entrance to the pipe, we can use Bernoulli's equation to get

$$p_1 + \frac{\rho v_1^2}{2} + \rho g h_1 = p_2 + \frac{\rho v_2^2}{2} + \rho g h_2$$
 (35)

With  $v_1 = 0$  and  $h_1 - h_2 = h$ , we get

$$p_2 - p_1 = \rho g h - \frac{\rho v_2^2}{2}$$
(36)

If we use a sufficiently long and small diameter pipe, the pipe flow velocity will be sufficiently small that we can neglect  $v_2^2$  compared to gh. Noting that both  $p_3$  and  $p_1$  are both atmospheric pressure and thus equal, we get for the pressure difference  $(p_2-p_3)$ at the ends of the pipe



Apparatus to measure the viscosity coefficient.

The pressure force on the fluid at the front end of the pipe is  $p_2A_2 = p_2A$  where A is the cross sectional area of the pipe. At the far end it is  $-p_3A$ , the minus sign is used because the pressure force is in the -z direction. Thus the net pressure force  $\vec{F}_p$  is

$$\vec{F}_{p} = \hat{z}(p_{2}A - p_{3}A)$$

$$= \hat{z}(p_{2} - p_{3})A \qquad (38)$$

$$= \hat{z}\rho ghA$$

If we divide  $\vec{F}_p$  by the volume AL of the pipe, we get the average pressure force per unit volume  $\vec{f}_p$ .

$$\vec{f}_{p} = \frac{\vec{F}_{p}}{AL} = \frac{\hat{z}}{AL}\rho ghA$$
  
$$\vec{f}_{p} = \hat{z}\frac{\rho gh}{L} \qquad \begin{array}{l} average \ pressure \\ force \ per \\ unit \ area \end{array}$$
(39)

As we mentioned, for steady laminar flow, the viscous force should be exactly opposed by the pressure force so that there is no acceleration of the fluid. Since the viscous force per unit volume is uniform throughout the fluid for parabolic pipe flow, the pressure force per unit volume should also be uniform, with the result that Equation (39) for  $\vec{f}_P$  should apply at all points in the fluid in the pipe. (There will always be some disturbance at the beginning of the flow that we are neglecting.)

Saying that the viscous and pressure forces oppose each other throughout the pipe flow gives us from Equations (34) for  $\vec{f}_v$  and (39) for  $\vec{f}_p$ 

$$\vec{f}_{p} = -\vec{f}_{v}$$

$$\hat{z} \frac{\rho g h}{L} = -\left[-\hat{z} \frac{4V_{0}}{R^{2}}\mu\right]$$

$$\frac{\rho g h}{L} = \frac{4V_{0}\mu}{R^{2}}$$
(40)

We are left with an equation involving measurable constants and the viscosity coefficient  $\mu$ .

Later in the text, we will see that the ratio  $\mu/\rho$ , which is called the *kinematic viscosity* coefficient v, is more convenient for theoretical work. Equation (40) gives us for this ratio

$$\nu \equiv \frac{\mu}{\rho} = \frac{gh}{L} \times \frac{R^2}{4V_0}$$
 kinematic viscosity  
determined from  
parabolic pipe (41)  
flow

The only constant that may be a bit difficult to measure directly is the stream velocity  $V_0$  at the center. This can be accurately determined by measuring the flow rate which we will call  $\Phi$  (phi), and then express  $V_0$  in terms of  $\Phi$ . We have called the flow rate  $\Phi$  because it is simply the flux  $\Phi$  of the fluid through the pipe, given by our old flux formula

$$\Phi = \int_{\substack{\text{area of}\\\text{tube}}} \vec{v} \cdot d\vec{A}$$
(42)

and is measured, in the MKS system, in cubic meters per second.

To calculate  $\Phi$ , we divide the cross sectional area into circular bands of radius r, thickness dr, as shown in Figure (7). The area of a band is  $2\pi r dr$  and the flux  $d\Phi$  through the band is

$$d\Phi = 2\pi r v(r) dr \tag{43}$$

With v(r) given by the parabolic profile  $(V_0/R^2)(R^2 - r^2)$ , we get for the total flux

$$\Phi = \int_{0}^{R} d\Phi = \int_{0}^{R} \frac{V_{0}}{R^{2}} (R^{2} - r^{2}) 2\pi r dr$$

$$= 2\pi V_{0} \left[ \int_{0}^{R} r dr - \frac{1}{R^{2}} \int_{0}^{R} r^{3} dr \right]$$

$$= 2\pi V_{0} \left[ \frac{r^{2}}{2} \Big|_{0}^{R} - \frac{1}{R^{2}} \frac{r^{4}}{4} \Big|_{0}^{R} \right]$$

$$\Phi = 2\pi V_{0} \left[ \frac{R^{2}}{2} - \frac{R^{2}}{4} \right] = \frac{V_{0}}{2} \pi R^{2} \qquad (44)$$

Since  $V_0(\pi R^2)$  is the flux we would get if the velocity were a uniform  $V_0$  across the pipe, we see that the flow rate for a parabolic profile is half that for a uniform flow.

With Equations (41) and (44) we can now express the kinematic viscosity  $\nu$  in terms of the easily measured volume flux  $\Phi$ . From Equation (44) we get



The integration area is the area  $2\pi R$  dr of the band.

$$V_0 = \frac{2\Phi}{\pi R^2}$$
 ;  $\frac{1}{V_0} = \frac{\pi R^2}{2\Phi}$ 

and from Equation (41) we get

$$\nu = \frac{\mu}{\rho} = \frac{gh}{L} \frac{R^2}{4} \frac{1}{V_0}$$
$$= \frac{gh}{L} \frac{R^2}{4} \frac{\pi R^2}{2\Phi}$$

$$v = \frac{\pi g h R^4}{8L \Phi} \begin{cases} formula for \\ measuring \\ kinematic \\ viscosity \end{cases}$$
(45)

Although rather a mess of constants appears in our formula for the kinematic viscosity  $\nu$ , all are quite easily measured. Note that by going to the kinematic viscosity, the result is independent of the density of the fluid.

#### **Exercise 1**

Show that the kinematic viscosity  $\nu$  has the dimensions of meters  $^2/\text{second}$  .

The two fluids that we will most often use in any discussion of fluid dynamics are water and air. At room temperature and pressure, the kinematic viscosity v of these two fluids are approximately

$$v_{\text{water}} = 1.0 \times 10^{-6} \text{meter}^2/\text{second}$$
  
 $v_{\text{air}} = 1.5 \times 10^{-5} \text{meter}^2/\text{second}$  (46)

Intuitively you would think that air would be much less viscous than water, but the two coefficients  $v_{air}$  and  $v_{water}$  are quite close, with air having the greater value. What has happened is that we have divided by the density, which brings the viscosity coefficients much closer together.

# Appendix: The Operator $\nabla^2$ in Spherical Polar Coordinates

# SPHERICAL POLAR COORDINATES

We will begin with a review of spherical coordinates discussed in Chapter 3. In spherical polar coordinates, the three unit vectors are  $\hat{r}$ ,  $\hat{\theta}$ ,  $\hat{\phi}$  are shown in Figure (A1) which is Figure (3-10) repeated. We have a complication in evaluating  $\nabla^2 f$  in spherical polar coordinates because these unit vectors change direction as we move about, and we can no longer set the derivatives of the unit vectors to zero. Thus we have to evaluate derivatives of the unit vectors as well as use the rather messy formula for  $\nabla f$  we derived in Equation (3-88)

$$\vec{\nabla}f(\mathbf{r},\theta,\phi) = \hat{\mathbf{r}}\frac{\partial f}{\partial \mathbf{r}} + \frac{\hat{\theta}}{\mathbf{r}}\frac{\partial f}{\partial \theta} + \frac{\hat{\phi}}{\mathbf{r}}\frac{\partial f}{\partial \phi} - (3-88)$$

What we have to evaluate is the complete expression

$$\nabla^{2} f(r,\theta,\phi) \equiv \nabla \cdot \nabla f(r,\theta,\phi)$$

$$= \left( \hat{r} \frac{\partial}{\partial r} + \frac{\hat{\theta}}{r} \frac{\partial}{\partial \theta} + \frac{\hat{\phi}}{r \sin \theta} \frac{\partial}{\partial \phi} \right)$$

$$\cdot \left( \hat{r} \frac{\partial f}{\partial r} + \frac{\hat{\theta}}{r} \frac{\partial f}{\partial \theta} + \frac{\hat{\phi}}{r \sin \theta} \frac{\partial f}{\partial \phi} \right)$$
(A1)

 $\begin{array}{c}
 z \\
 \theta \\
 \overline{r} \\
 \overline{r}$ 

**Figure A1 (3-10 repeated)** Unit vectors in spherical polar coordinates.

This product involves terms like

$$\begin{pmatrix} \hat{\theta} \\ \hat{r} \\ \frac{\partial}{\partial \theta} \end{pmatrix} \cdot \left( \hat{r} \frac{\partial f}{\partial r} \right) = \frac{\hat{\theta}}{r} \cdot \left[ \frac{\partial \hat{r}}{\partial \theta} \frac{\partial f}{\partial r} + \hat{r} \frac{\partial^2 f}{\partial \theta \partial r} \right]$$

$$= \frac{1}{r} \left( \hat{\theta} \cdot \frac{\partial \hat{r}}{\partial \theta} \right) \frac{\partial f}{\partial r} + \left( \hat{\theta} \cdot \hat{r} \right) \frac{1}{r} \frac{\partial^2 f}{\partial \theta \partial r}$$
(A2)

Because the unit vectors always remain perpendicular to each other as we move around in space,  $\hat{\theta} \cdot \hat{r} = 0$  and the second term in Equation (A2) is zero. However, when we change the angle  $\theta$ , the unit vector  $\hat{r}$  changes direction. For example, at  $\theta = 0$ ,  $\hat{r}$  points straight up, but at  $\theta = 90^{\circ}$ ,  $\hat{r}$  is horizontal. Thus  $\partial \hat{r} / \partial \theta$  is not zero and has to be evaluated.

In order to evaluate Equation (A1) for  $\nabla^2 f$ , we will first calculate all the derivatives of all the unit vectors, and then plug the whole mess together. We find derivatives like  $\partial \hat{r} / \partial \theta$  by evaluating the change  $\Delta \hat{r}$  as we make a change  $\Delta \theta$  and then taking the limit as  $\Delta \theta$  goes to zero. The nine derivatives are as follows.

## Derivatives of r

#### 1) Change of $\hat{r}$ with r

$$\frac{\partial \hat{\mathbf{f}}}{\partial \mathbf{r}} = \mathbf{0} \qquad \begin{array}{c} because \ \hat{r} \ does \ not \\ change \ direction \ as \ we \\ go \ out \ along \ a \ radius \end{array}$$

#### 2) Change of $\hat{\mathbf{r}}$ with $\theta$

Figure (A2) shows  $\Delta \vec{r}$  that we get when  $\theta$  increases by  $\Delta \theta$ . We see that  $\overrightarrow{\Delta r}$  points in the  $\hat{\theta}$  direction and has a length  $|\vec{\Delta r}| = |\hat{r}| \Delta \theta = 1 \times \Delta \theta$ . Thus we get





 $\overrightarrow{\Delta r}$  points in the  $\hat{\theta}$  direction



X

#### 3) Change of $\hat{\mathbf{r}}$ with $\phi$

In Figure (A3), when we go from  $\phi$  to  $\phi + \Delta \phi$ , the unit vector  $\hat{\mathbf{r}}$  goes to the unit vector  $\hat{\mathbf{r}}'$ . The projections of  $\hat{r}$  and  $\hat{r}'$  in the horizontal plane have a length  $\hat{r} | \sin \phi = 1 \sin \phi$ , and differ in direction by an angle  $\Delta \phi$ . The change  $\Delta \vec{r} = \hat{r}' - \hat{r}$  points in the  $\hat{\phi}$  direction, and has the same length as the change in the horizontal projections of  $\hat{r}$  and  $\hat{r}'$ , which from the small triangle is seen to be  $(\sin\phi)(\Delta\phi)$ . Thus

$$\vec{\Delta \mathbf{r}} = \hat{\phi}(\sin\phi)\Delta\phi$$
$$\frac{\vec{\Delta r}}{\Delta\phi} = \boxed{\hat{\phi}(\sin\phi) = \frac{\partial \hat{r}}{\partial\phi}}$$
(A4)



 $\Delta \hat{\mathbf{r}}$  is the change in the unit vector  $\hat{\mathbf{r}}$ when we increase  $\phi$  by  $\Delta \phi$ .



Unit vectors enlarged.

Figure A3 Evaluation of  $\partial \hat{r} / \partial \phi$ .

# Derivatives of $\hat{\theta}$

# 4) Change of $\hat{\theta}$ with r

None of the unit vectors change direction as we go out along the radius, thus

$$\frac{\partial \hat{\theta}}{\partial r} = 0 \tag{A5}$$

## 5) Change of $\hat{\theta}$ with $\theta$

From Figure (A4) we see that as we increase  $\theta$  to  $\theta + \Delta \theta$ , the unit vector  $\hat{\theta}$  goes to  $\hat{\theta}' = \hat{\theta} + \vec{\Delta \theta}$ . From the small triangle, we see that the change  $\vec{\Delta \theta}$  points in the  $-\hat{r}$  direction, and has a magnitude  $\Delta \theta$ . Thus we have

$$\vec{\Delta \theta} = (-\hat{r}) \Delta \theta$$
$$\vec{\Delta \theta} = \boxed{-\hat{r} = \frac{\partial \hat{\theta}}{\partial \theta}}$$
(A6)



 $\overrightarrow{\Delta \theta}$  is the change in the unit vector  $\hat{\theta}$  when we increase  $\theta$  by  $\Delta \theta$ .



 $\overrightarrow{\Delta \theta}$  points in the  $-\hat{r}$  direction



## 6) Change of $\hat{\theta}$ with $\phi$

From Figure (A5), we see that  $\hat{\theta}$  changes to  $\hat{\theta}'$  as  $\phi$ goes to  $\phi + \Delta \phi$ . The change  $\Delta \vec{\theta}$  points in the  $\hat{\phi}$ direction. To determine the magnitude of  $\Delta \vec{\theta}$ , note that  $\Delta \vec{\theta}$  and its projection in the horizontal plane are the same. Since the projections of  $\hat{\theta}$  and  $\hat{\theta}'$  have a length of  $\cos \theta$ , and an angle  $\Delta \phi$  between them, the length of  $\Delta \vec{\theta}$  is  $\cos \theta \Delta \phi$  as seen in the small horizontal triangle. Thus

$$\vec{\Delta \theta} = \hat{\phi}(\cos \phi) \Delta \phi$$
$$\vec{\Delta \theta} = \left[ \hat{\phi} \cos \theta = \frac{\partial \hat{\theta}}{\partial \phi} \right]$$
(A7)

# Derivatives of $\hat{\phi}$

## 7) Change of $\hat{\phi}$ with r

As we noted earlier, the unit vectors do not change with r, thus



 $\overrightarrow{\Delta \theta}$  is the change in the unit vector  $\hat{\theta}$  when we increase  $\phi$  by  $\Delta \phi$ .



 $\overrightarrow{\Delta \theta}$  points in the  $\hat{\phi}$  direction

**Figure A5** *Evaluation of*  $\partial \hat{\theta} / \partial \phi$ .

#### 8) Change of $\hat{\phi}$ with $\theta$

As we can see from Figure (A6), the unit vector  $\hat{\phi}$  does not change direction when we change the angle  $\theta$ .

For example, when  $\vec{r}$  is in the xz plane,  $\hat{\phi}$  points in the  $+\hat{y}$  direction for all angles  $\theta$ . Thus

$$\frac{\partial \hat{\phi}}{\partial \theta} = 0 \tag{A9}$$

## 9) Change of $\hat{\phi}$ with $\phi$

Finally, we have to figure out how the unit vector  $\hat{\phi}$  changes with the angle  $\phi$ . This time we will take a top down view as shown in Figure (A7). When we change  $\phi$  to  $\phi + \Delta \phi$ , the unit vector  $\hat{\phi}$  goes to  $\hat{\phi}'$ . From the small triangle we see that the change  $\Delta \hat{\phi}$  points *toward* the  $\hat{z}$  axis and has a magnitude  $\Delta \phi$ .

In Figure (A8), we see that a unit vector  $\hat{u}$  pointing toward the z axis is given by

 $\left. \begin{array}{c} \text{unit vector} \\ \text{pointing} \\ \text{toward} \\ \text{z axis} \end{array} \right\} = -\hat{r}\sin\theta - \hat{\theta}\cos\theta$ 

Thus  $\Delta \hat{\phi} = \Delta \phi (-\hat{r} \sin \theta - \hat{\theta} \cos \theta)$  and we get



The unit vector  $\hat{\phi}$  does not change when we increase  $\theta$  by  $\Delta \theta$ .

#### **Figure A6**

Evaluation of  $\partial \hat{\phi} / \partial \theta$ .







Unit vectors enlarged. The unit vector  $-\hat{u}$  points toward the z axis.

#### **Figure A7** *Evaluation of* $\partial \hat{\theta} / \partial \phi$ .





# **Summary of Derivatives of Unit Vectors**

In summary, we get

$$\frac{\partial \hat{\mathbf{r}}}{\partial \mathbf{r}} = 0; \quad \frac{\partial \hat{\mathbf{r}}}{\partial \theta} = \hat{\theta}; \quad \frac{\partial \hat{\mathbf{r}}}{\partial \phi} = \hat{\phi} \sin \theta$$
$$\frac{\partial \hat{\theta}}{\partial \mathbf{r}} = 0; \quad \frac{\partial \hat{\theta}}{\partial \theta} = -\hat{\mathbf{r}}; \quad \frac{\partial \hat{\theta}}{\partial \phi} = \hat{\theta} \cos \theta$$
$$\frac{\partial \hat{\phi}}{\partial \mathbf{r}} = 0; \quad \frac{\partial \hat{\phi}}{\partial \theta} = 0; \quad \frac{\partial \hat{\phi}}{\partial \phi} = -\hat{\mathbf{r}} \sin \theta - \hat{\theta} \cos \theta$$
(A11)

# Calculation of $\nabla^2 f$

We are now ready to calculate  $\nabla^2 f$  given again by Equation (A1)

 $\rightarrow \rightarrow$ 

$$\begin{split} \nabla^{2} f(\mathbf{r}, \theta, \phi) &= \vec{\nabla} \cdot \vec{\nabla} f(\mathbf{r}, \theta, \phi) \\ &= \left( \hat{\mathbf{r}} \frac{\partial}{\partial \mathbf{r}} + \frac{\hat{\theta}}{\mathbf{r}} \frac{\partial}{\partial \theta} + \frac{\hat{\phi}}{\mathbf{r} \sin \theta} \frac{\partial}{\partial \phi} \right) \\ \cdot \left( \hat{\mathbf{r}} \frac{\partial \mathbf{f}}{\partial \mathbf{r}} + \frac{\hat{\theta}}{\mathbf{r}} \frac{\partial \mathbf{f}}{\partial \theta} + \frac{\hat{\phi}}{\mathbf{r} \sin \theta} \frac{\partial \mathbf{f}}{\partial \phi} \right) \\ &= \hat{\mathbf{r}} \cdot \left[ \frac{\partial \hat{\mathbf{r}}}{\partial \mathbf{r} \partial \mathbf{r}} + \hat{\mathbf{r}} \frac{\partial^{2} \mathbf{f}}{\partial \mathbf{r}^{2}} + \frac{\partial \hat{\theta}}{\partial \mathbf{r}} \frac{1}{\mathbf{r}} \frac{\partial \mathbf{f}}{\partial \theta} + \hat{\theta} \frac{\partial}{\partial \mathbf{r}} \left( \frac{1}{\mathbf{r} \partial \theta} \right) \right] \\ &+ \frac{\partial \hat{\phi}}{\partial \mathbf{r}} \frac{1}{\mathbf{r} \sin \theta} \frac{\partial \mathbf{f}}{\partial \phi} + \hat{\phi} \frac{\partial}{\partial \mathbf{r}} \left( \frac{1}{\mathbf{r} \sin \theta} \frac{\partial \mathbf{f}}{\partial \phi} \right) \right] \\ &+ \frac{\hat{\theta}}{\mathbf{r}} \cdot \left[ \frac{\partial \hat{\mathbf{f}}}{\partial \theta} \frac{\partial \mathbf{f}}{\partial \mathbf{r}} + \hat{\mathbf{r}} \frac{\partial^{2} \mathbf{f}}{\partial \theta \partial \mathbf{r}} + \frac{\partial \hat{\theta}}{\partial \theta} \frac{1}{\mathbf{r} \partial \theta} + \frac{\hat{\theta}}{\theta} \frac{\partial^{2} \mathbf{f}}{\partial \theta^{2}} \right] \\ &+ \frac{\partial \hat{\phi}}{\partial \theta} \frac{1}{\mathbf{r} \sin \theta} \frac{\partial \mathbf{f}}{\partial \phi} + \hat{\phi} \frac{\partial}{\partial \theta} \left( \frac{1}{\mathbf{r} \sin \theta} \frac{\partial \mathbf{f}}{\partial \phi} \right) \right] \\ &+ \frac{\hat{\phi}}{\mathbf{r} \sin \theta} \cdot \left[ \frac{\partial \hat{\mathbf{r}}}{\partial \phi} \frac{\partial \mathbf{f}}{\partial \mathbf{r}} + \hat{\mathbf{r}} \frac{\partial^{2} \mathbf{f}}{\partial \phi \partial \mathbf{r}} + \frac{\partial \hat{\theta}}{\partial \phi} \frac{1}{\mathbf{r}} \frac{\partial \mathbf{f}}{\partial \phi} \right] \\ &+ \frac{\hat{\theta}}{\mathbf{r} \sin \theta} \cdot \left[ \frac{\partial \hat{\mathbf{r}}}{\partial \phi} \frac{\partial \mathbf{f}}{\partial \mathbf{r}} + \hat{\mathbf{r}} \frac{\partial^{2} \mathbf{f}}{\partial \phi \partial \mathbf{r}} \right] \\ &+ \hat{\theta} \left( \frac{1}{\mathbf{r} \sin \theta} \frac{\partial \mathbf{f}}{\partial \phi} \right] \end{aligned} \tag{A12}$$

The terms in Equation (58) with a single line through them are zero because the unit vectors are orthogonal: i.e.,  $\hat{\mathbf{r}} \cdot \hat{\mathbf{\theta}} = 0$ ,  $\hat{\mathbf{\theta}} \cdot \hat{\mathbf{\phi}} = 0$ , etc. Next we use our summary, Equation (57) to evaluate the following terms.

$$\hat{\mathbf{r}} \cdot \frac{\partial \hat{\mathbf{r}}}{\partial \mathbf{r}} = 0$$
 because  $\frac{\partial \hat{\mathbf{r}}}{\partial \mathbf{r}} = 0$  (A13a)

$$\hat{\mathbf{r}} \cdot \frac{\partial \hat{\theta}}{\partial \mathbf{r}} = 0$$
 because  $\frac{\partial \hat{\theta}}{\partial \mathbf{r}} = 0$  (A13b)

$$\hat{\mathbf{r}} \cdot \frac{\partial \hat{\phi}}{\partial \mathbf{r}} = 0$$
 because  $\frac{\partial \hat{\phi}}{\partial \mathbf{r}} = 0$  (A13c)

$$\hat{\theta} \cdot \frac{\partial \hat{r}}{\partial \theta} = \hat{\theta} \cdot \hat{\theta} = 1$$
 (A13d)

$$\hat{\theta} \cdot \frac{\partial \hat{\theta}}{\partial \theta} = \hat{\theta} \cdot (-\hat{r}) = 0$$
 (A13e)

$$\hat{\theta} \cdot \frac{\partial \hat{\phi}}{\partial \theta} = 0$$
 because  $\frac{\partial \hat{\phi}}{\partial \theta} = 0$  (A13f)

$$\hat{\phi} \cdot \frac{\partial \hat{r}}{\partial \phi} = \hat{\phi} \cdot (\hat{\phi} \sin \theta) = \sin \theta$$
 (A13g)

$$\hat{\phi} \cdot \frac{\partial \hat{\theta}}{\partial \phi} = \hat{\phi} \cdot (\hat{\phi} \cos \theta) = \cos \theta$$
 (A13h)

$$\hat{\phi} \cdot \frac{\partial \hat{\phi}}{\partial \phi} = \hat{\phi} \cdot (-\hat{r} \sin \theta - \hat{\theta} \cos \theta) = 0$$
 (A13i)

The terms in Equation (A12) for  $\nabla^2 f$ , that are zero because of Equations (A13), have a double line through them. We are left with

$$\nabla^{2} \mathbf{f} = \hat{\mathbf{r}} \cdot \hat{\mathbf{r}} \frac{\partial^{2} \mathbf{f}}{\partial r^{2}} + \frac{\hat{\theta}}{\mathbf{r}} \cdot \frac{\partial \hat{\mathbf{r}}}{\partial \theta} \frac{\partial \mathbf{f}}{\partial \mathbf{r}}$$
$$+ \frac{\hat{\theta}}{\mathbf{r}} \cdot \frac{\hat{\theta}}{\mathbf{r}} \frac{\partial^{2} \mathbf{f}}{\partial \theta^{2}} + \frac{\hat{\phi}}{\mathbf{r} \sin \theta} \cdot \frac{\partial \hat{\mathbf{r}}}{\partial \phi} \frac{\partial \mathbf{f}}{\partial \mathbf{r}}$$
$$+ \frac{\hat{\phi}}{\mathbf{r} \sin \theta} \cdot \frac{\partial \hat{\theta}}{\partial \phi} \frac{1}{\mathbf{r}} \frac{\partial \mathbf{f}}{\partial \theta} + \frac{\hat{\phi}}{\mathbf{r} \sin \theta} \cdot \frac{\hat{\phi}}{\mathbf{r} \sin \theta} \frac{\partial^{2} \mathbf{f}}{\partial \phi^{2}}$$
(A14)

Using Equations (A13), Equation (A14) becomes

$$\nabla^{2} f = \frac{\partial^{2} f}{\partial r^{2}} + \frac{1}{r} \frac{\partial f}{\partial r} + \frac{1}{r^{2}} \frac{\partial^{2} f}{\partial \theta^{2}} + \frac{1}{r \sin \theta} \sin \theta \frac{\partial f}{\partial r} + \frac{1}{r \sin \theta} \cos \theta \frac{1}{r} \frac{\partial f}{\partial \theta} + \frac{1}{r^{2} \sin^{2} \theta} \frac{\partial^{2} f}{\partial \phi^{2}}$$
(A15)

This becomes

$$\nabla^{2} \mathbf{f} = \frac{\partial^{2} \mathbf{f}}{\partial \mathbf{r}^{2}} + \frac{2}{\mathbf{r}} \frac{\partial \mathbf{f}}{\partial \mathbf{r}} + \frac{1}{\mathbf{r}^{2}} \left( \frac{\cos \theta}{\sin \theta} \frac{\partial \mathbf{f}}{\partial \theta} + \frac{\partial^{2} \mathbf{f}}{\partial \theta^{2}} \right) + \frac{1}{\mathbf{r}^{2} \sin^{2} \theta} \frac{\partial^{2} \mathbf{f}}{\partial \phi^{2}}$$
(A16)

In most textbooks, you will find the equivalent formula

$$\nabla^{2} \mathbf{f} = \frac{1}{r} \frac{\partial^{2}}{\partial r^{2}} (\mathbf{r} \mathbf{f}) + \frac{1}{r^{2}} \left[ \frac{1}{\sin \theta} \frac{\partial}{\partial \theta} \left( \sin \theta \frac{\partial \mathbf{f}}{\partial \theta} \right) + \frac{1}{\sin^{2} \theta} \frac{\partial^{2} \mathbf{f}}{\partial \phi^{2}} \right]$$
(10)

which is the result we stated earlier in the chapter.

#### **Exercise 2**

Show that Equation (A15) follows from Equations (A13) and (A14).

#### **Exercise 3**

Show that Equations (A16) and (10) are equivalent.

# Calculus 2000-Chapter 5 Introduction to Complex Variables

# A ROAD MAP

In this chapter you will see that the use of complex variables greatly simplifies the analysis of RLC circuits and other forms of sinusoidal behavior. This chapter does not depend on previous chapters of the Calculus text and may be studied directly in connection with the related material in Chapters 27 and 31 of the Physics text.

This chapter is also background material for the next chapter, Chapter 6, on the Schrödinger's equation. The wave equations we have discussed so far can be solved using either real variables or complex variables. Schrödinger's wave equation is different in that the equation itself involves complex numbers and cannot be handled by real variables alone. That is why this chapter is a prerequisite. Also, to solve Schrödinger's equation for the hydrogen atom requires the use of  $\nabla^2$  in spherical polar coordinates, which we discussed in the last chapter.

Once we finish Chapters 5 and 6 on complex variables and Schrödinger's equation, we return to basic calculus operations, discussing divergence in Chapter 7 and curl in Chapter 8. We then apply divergence and curl to electromagnetism in Chapters 9, 10, and 11, and to fluid dynamics in Chapters 12 and 13.

## INTRODUCTION

After introducing the concepts of imaginary and complex numbers, we find that an important feature of a complex number is that it can be expressed as a complex exponential. We then go on to two major applications of complex variables that we just mentioned. One is the analysis of RLC circuits, which can be handled using real variables only, but where there is an enormous simplification if we use complex variables. Then in the next chapter, we discuss the Schrödinger's equation where the equation itself involves complex variables.

There are other topics involving the theory of complex variables that we will not discuss in these introductory chapters. It is possible to construct fascinating maps of complex functions and distort these maps in intriguing ways (not completely unlike the distortion of images one can create on the computer). Complex variables are also useful in finding the formulas for various integrals. These advanced topics are usually covered in a graduate level mathematical physics course.

#### **IMAGINARY NUMBERS**

What number, when multiplied by itself gives (-1)? The answer is none of the ordinary numbers. This number,  $\sqrt{-1}$  is not one of the real numbers like 5, -2, 3, etc. It belongs to a completely different system of numbers which we call *imaginary* numbers. The number  $\sqrt{-1}$  is denoted by the letter *i*, and the square root of any negative number can be written as a real number times *i*. For example

$$\sqrt{-7} = \sqrt{7 \times (-1)} = \sqrt{7} \times \sqrt{-1}$$

$$= (\sqrt{7})i$$

$$example 
of an 
imaginary (1) 
number$$

All numbers with one factor of *i* are imaginary.

# **COMPLEX NUMBERS**

We can make things a bit more complicated by adding together a real number and an imaginary number, such as (4 + 3i). Such a mixture with both a *real part* (4) and an *imaginary part* (3i) is called a *complex number*. These two parts are distinct; there is no way we can confuse the real and imaginary parts because imaginary numbers are not part of the real number system.

This is not the first time we have encountered a quantity that has two distinct parts. In our strobe



photographs showing the motion of a ball, we noted that the position of the ball could be described by the *coordinate vector*  $\vec{r}$ , as shown in Figure (1). For the strobe photographs, which only show two dimensions, the coordinate vector  $\vec{r}$  was completely specified by its (x) and (y) components. Thus two dimensional coordinate vectors and complex numbers are similar in that they both consist of two independent components.

This similarity suggests that we can treat a complex number in the same way we handle a two dimensional coordinate vector, plotting the real and imaginary parts along different axes. It is traditional to plot the real part along the x axis and the imaginary part along the y axis. Thus, for example, the complex number (4 + 3i) can be represented by a point whose coordinate vector has an x component of 4 and a y component of 3 as shown in Figure (2).

In this chapter you will see that in some cases there is considerable simplification of the mathematics and much greater insight when we use complex numbers. This is illustrated in our analysis of the RLC circuit where we will see that a sinusoidal oscillation and an exponential decay can both be handled by one simple complex function.



Plot of the complex number (4 + 3i), where the real part is plotted along the x axis and the imaginary part long the y axis.

# EXPONENTIAL FORM OF THE COMPLEX NUMBER

Once we start plotting complex numbers on x and y axes, we will find that any complex number can be expressed in the exponential form  $re^{i\theta}$ . How we get to this rather remarkable result can be seen in the following way.

Let us go back to Figure (2), showing the plot of the complex number (4 + 3i). One way to describe that point is to give its x and y coordinates (x = 4, y = 3i). An equally good description, shown in Figure (2a), is to give the distance r from the origin to the point, and the angle  $\theta$  that  $\vec{r}$  makes with the x or real axis. From the Pythagorean theorem we have

$$r^{2} = x^{2} + y^{2} = 4^{2} + 3^{2} = 16 + 9 = 25$$
  
r = 5 (2)

The tangent of the angle  $\theta$  is the opposite side y divided by the adjacent side x

$$\tan \theta = \frac{y}{x} = \frac{3}{4} = .75$$
 (3)

Entering .75 in our calculator and pressing the tan<sup>-1</sup> button gives 36.9°. Thus the point is located at a distance r = 5 from the origin at an angle  $\theta = 36.9^\circ$ .

It is traditional to use the letter (z) to describe a complex number. Thus if a complex number (z) has a real part (x) and an imaginary part (iy), we can write

$$z = x + iy \tag{4}$$



Plot of the complex number (4 + 3i), showing the angle  $\theta$ .

Now let us express z in terms of the variables r and  $\theta$  rather than x and y. Since from Figure (2a) we see that

we can write (z) as

$$z = x + iy = r\cos\theta + i r\sin\theta$$
$$z = r(\cos\theta + i \sin\theta)$$
(6)

It is the function  $(\cos \theta + i \sin \theta)$  that we wish to study in detail.

Let us first look at the derivative of  $(\cos \theta + i \sin \theta)$ . Since

$$\frac{d}{d\theta}\cos\theta = -\sin\theta; \quad \frac{d}{d\theta}\sin\theta = \cos\theta \qquad (7)$$

we get

$$\frac{\mathrm{d}}{\mathrm{d}\theta}(\cos\theta + i\sin\theta) = -\sin\theta + i\cos\theta$$

Since  $(-1) = i^2$ , this can be written

$$\frac{d}{d\theta}(\cos\theta + i\sin\theta) = i^2\sin\theta + i\cos\theta$$

$$= i(\cos\theta + i\sin\theta)$$
(8)

To express this result more formally, let us write

$$f(\theta) = (\cos \theta + i \sin \theta)$$
(9)

Then Equation (8) becomes

$$\frac{\mathrm{d}}{\mathrm{d}\theta}\mathbf{f}(\theta) = i\mathbf{f}(\theta) \tag{8a}$$

To within a constant (*i*), the function  $f(\theta)$  is equal to its own derivative. What function that you are already familiar with, behaves this way? The *exponential function*! Recall that

$$\frac{\mathrm{d}}{\mathrm{d}x}\mathrm{e}^{\mathrm{a}x} = \mathrm{a}\mathrm{e}^{\mathrm{a}x} \tag{10}$$

Thus if we replace (x) by  $\theta$  and (a) by (i), we get

$$\frac{\mathrm{d}}{\mathrm{d}\theta}\mathrm{e}^{i\theta} = i\mathrm{e}^{i\theta} \tag{11}$$

Comparing Equations (8) and (11), we see that the function  $(\cos \theta + i \sin \theta)$  and the function  $e^{i\theta}$  obey the same rule for differentiation.

When two functions  $(\cos \theta + i \sin \theta)$  and  $e^{i\theta}$  have the same derivatives, does that mean that they are the same functions? It will if we show that both functions start off with the same value for small values of  $\theta$ . Then as we increase  $\theta$ , if both functions have the same derivative or slope, they must continue to be the same function for all values of  $\theta$ .

#### **Small Angle Approximations**

We can show that  $(\cos \theta + i \sin \theta)$  and  $e^{i\theta}$  have the same values for small  $\theta$  by using the small angle approximations for  $\sin \theta$ ,  $\cos \theta$  and  $e^{i\theta}$ . In our discussion of the exponential in Chapter 1 of the Calculus text, (Cal 1, Eq. 136), we had

$$e^{x} = 1 + x + \frac{x^{2}}{2!} + \frac{x^{3}}{3!} + \cdots$$
 (1-136)

While this expansion is true for any value of x, it is most useful for small values of x where we do not have to keep many terms to get an accurate answer.

Setting  $x = i\theta$  gives

$$e^{i\theta} = 1 + i\theta + \frac{i^2\theta^2}{2!} + \frac{i^3\theta^3}{3!} + \cdots$$
 (12)

(Since our previous discussion of exponents only dealt with real numbers, we can consider Equation (12) as the definition of what we mean when the exponent is a complex number).

What we did not discuss earlier were the expansions for  $\cos\theta$  and  $\sin\theta$ . Let us state them and check their accuracy now. They are

$$\cos \theta = 1 - \frac{\theta^2}{2!} + \frac{\theta^4}{4!} + \cdots$$
 (13)

$$\sin \theta = \theta - \frac{\theta^3}{3!} + \frac{\theta^5}{5!} + \cdots$$
 (14)

where  $\theta$  is in radians. Again these expansions are valid for any value of  $\theta$ , but most useful for small values where we do not have to keep many terms.

Let us check the accuracy of these expansions for  $\theta = .1$  radians. We have, keeping three terms,

$$\cos (.1) = 1 - \frac{(.1)^2}{2!} + \frac{(.1)^4}{4!}$$
  
= 1 -  $\frac{.01}{2} + \frac{.0001}{4 \times 3 \times 2}$   
= 1 - .005 + .0000004166 (15)  
= .995004166

Changing our calculator from degrees to radians and taking the cos(.1) gives

$$\cos(.1) = .995004165$$
 (16)

We see that we get almost a nine place accuracy by keeping the first three terms of the expansion.

For sin(.1), keeping the first three terms, we have

$$\sin (.1) = .1 - \frac{(.1)^3}{3!} + \frac{(.1)^5}{5!}$$
$$= .1 - \frac{.001}{3 \times 2} + \frac{.0001}{5 \times 4 \times 3 \times 2}$$
$$= .1 - .0001666666 + .000000083$$
$$= .009833417$$
(17)

The calculator gives

$$\sin(.1) = .099833416 \tag{18}$$

which is again accurate to almost nine places.

If you can't figure out how to get your calculator to work in radians, you can convert .1 radians to degrees by using the conversion factor

$$\frac{360 \text{ degrees/cycle}}{2\pi \text{ radians/cycle}} = 57.29577951 \frac{\text{degrees}}{\text{radian}} (19)$$

Now that we have checked the expansions (13) and (14) let us see what the expansion for  $(\cos \theta + i \sin \theta)$  is. We get, replacing – signs by  $i^2$ , and using  $i^4 = +1$ ,

$$\cos \theta = 1 - \frac{\theta^2}{2!} + \frac{\theta^4}{4!} = 1 + \frac{i^2 \theta^2}{2!} + \frac{i^4 \theta^4}{4!}$$
$$i \sin \theta = i \left( \theta - \frac{\theta^3}{3!} + \frac{\theta^5}{5!} \right) = i \theta + \frac{i^3 \theta^3}{3!} + \frac{i^5 \theta^5}{5!}$$
(20)

Adding these gives

$$(\cos \theta + i \sin \theta) = 1 + i\theta + \frac{i^2 \theta^2}{2!} + \frac{i^3 \theta^3}{3!} + \frac{i^4 \theta^4}{4!} + \frac{i^5 \theta^5}{5!} + \cdots$$
(21)

which is just the expansion we had for  $e^{i\theta}$  in Equation (12).

Comparing Equations (20) and (21), you can see that the expansions for  $\cos \theta$  and  $i \sin \theta$  fit together to produce the much more regular expansion of  $e^{i\theta}$ . We will also see that it is often much easier to work with the complete function  $e^{i\theta}$  than with  $\cos \theta$  and  $\sin \theta$  separately.

In summary, if we define a complex exponential by the series expansion of Equation (12), then we have shown that

$$e^{i\theta} = \cos\theta + i\sin\theta \tag{22}$$

Even though we checked the sin and cosine expansions for a small value of  $\theta$ , the fact that  $e^{i\theta}$  and  $(\cos \theta + i \sin \theta)$  have the same derivative properties means that Equation (22) holds for all values of  $\theta$ .

If we replace  $\theta$  by  $-\theta$  in Equation (22) we get

$$e^{-i\theta} = \cos(-\theta) + i\sin(-\theta)$$

Since  $\cos(-\theta) = \cos\theta$ ,  $\sin(-\theta) = -\sin\theta$ , this gives

$$e^{-i\theta} = \cos\theta - i\sin\theta \tag{23}$$

If we add Equations (22) and (23), the  $\sin\theta$  terms cancel, and we are have, after dividing through by 2,

$$\cos\theta = \frac{e^{i\theta} + e^{-i\theta}}{2}$$
(24)

Subtracting Equation (23) from (22) cancels the  $\cos \theta$  terms, leaving, after dividing through by 2i,

$$\sin\theta = \frac{e^{i\theta} - e^{-i\theta}}{2i}$$
(25)

If we note that

$$\frac{1}{i} = \frac{1}{i} \times \frac{i}{i} = \frac{i}{i^2} = \frac{i}{-1} = -i$$

we can write Equation (25) as

$$\sin \theta = (-i) \frac{e^{i\theta} - e^{-i\theta}}{2}$$
$$\sin \theta = i \frac{e^{-i\theta} - e^{i\theta}}{2}$$
(25a)

Equations (22) through (25) give a complete prescription of how to go back and forth between  $\cos \theta$ ,  $\sin \theta$ ,  $e^{i\theta}$  and  $e^{-i\theta}$ .

Finally returning to our complex function



as the other way of expressing a complex number, where r is the distance from the origin and  $\theta$  the angle the coordinate vector  $\vec{r}$  makes with the x or real axis.

## Exercise 1

(a) Construct a series expansion for  $e^{-i\theta}$ .

(b) Using the series expansions for  $e^{i\theta}$  and  $e^{-i\theta}$  in Equation (25a), show that you end up with the series expansion for  $sin(\theta)$ .

#### The Complex Conjugate Z\*

The *complex conjugate* of a complex number is defined as the number we get by replacing all factors of (i) by (-i) in the formula for the number. We generally denote the complex conjugate by placing an asterisk after the number. For example, if

$$z = x + iy$$

then

$$\mathbf{z}^* = \mathbf{x} - i\mathbf{y} \tag{26a}$$

If we start with

$$z = re^{i\theta}$$

then

$$z^* = r e^{-i\theta} \tag{26b}$$

The main reason for defining a complex conjugate is that the product of a complex number z with its complex conjugate  $z^*$  is *always a real positive number*, equal to the square of the distance r that the complex point is from the origin. For our two examples above, we have

$$z^{*}z = (x - iy)(x + iy)$$
  
=  $x^{2} - ixy + iyx - i^{2}y^{2}$   
=  $x^{2} + y^{2} = r^{2}$ 

and

$$z^*z = (re^{-i\theta})(re^{i\theta}) = r^2$$
(26c)



#### Figure 3

The RC circuit. When we walk around in the direction shown by the circular arrow, we go with  $V_C$  but against  $V_R$ , giving  $V_C-V_R$  as the sum of the voltage rises.

# DIFFERENTIAL EQUATIONS FOR R, L, C CIRCUITS

One of the most convenient uses of complex variables is in the analysis of electric circuits involving resistors, capacitors and inductors. We will see that using complex variables unifies the analysis and greatly simplifies the work involved.

# **The RC Circuit**

Let us begin with the RC circuit shown in Figure (3). If we charge up the capacitor to a voltage  $V_0$ , and close the switch, a current flows out of the capacitor through the resistor, and the voltage  $V_C$  on the capacitor decays exponentially.

The formulas for the capacitor voltage  $V_C$  and resistor voltage  $V_R$  are

$$V_{\rm C} = \frac{Q}{C}$$
;  $V_{\rm R} = iR$  (27)

where Q is the charge on the capacitor, C the capacitor's capacitance, (i) the current through the resistor and R the resistor's resistance in ohms. It is assumed that C and R are constants and that (i) is the rate at which charge Q is leaving the capacitor. That is,

$$i = -\frac{dQ}{dt}$$
(28a)

Setting the sum of the voltage rises around the circuit equal to zero (see Equation 27-41 in the Physics text) and using (28a) for i, gives us

$$V_{\rm C} - V_{\rm R} = 0$$
  
$$\frac{Q}{C} - iR = \frac{Q}{C} + \frac{dQ}{dt}R = 0$$
 (28b)

Dividing through by R, we get

$$\frac{\mathrm{d}Q}{\mathrm{d}t} + \frac{\mathrm{Q}}{\mathrm{RC}} = 0 \tag{29}$$

as the differential equation for the amount of charge Q remaining in the capacitor.
### An Aside on Labeling Voltages

To avoid worrying about minus signs like the i = -dQ/dt for the discharging capacitor, we will obtain the differential equations for our L, R, and C circuits by sketching the voltages when the rate of change of charge in the capacitor and change of the current in an inductor are all **positive**. If we had an **increasing** current running down through three cir-

cuit elements R, L, and C, all the voltages would point up as shown in Figure (4). The resistance voltage  $V_R$  is always directed opposite to the current. If the downward current is increasing, then the inductor opposes the increase and points upward. With a positive current flowing into the capacitor, the current is equal to +dQ/dt. If the capacitor started off with zero charge, then the upper plate is becoming positively charged by the positive current flowing into it.

Using these conventions for

the current and voltages, we



Figure 4 Direction of the voltages for an increasing downward current.

can construct an RC circuit from Figure (4) by pulling out the inductor and connecting the back side of the circuit as shown in Figure (5). Setting the sum of the voltages to zero around Figure (5) gives (walking around the circuit

$$V_{\rm R} + V_{\rm C} = 0$$
  
iR +  $\frac{Q}{C} = 0$  (30)

counterclockwise as shown by the circular arrow)

With i = +dQ/dt, this gives

$$R\frac{dQ}{dt} + \frac{Q}{C} = 0 ; \qquad \frac{dQ}{dt} + \frac{Q}{RC} = 0$$
 (29a)

This is just the same as Equation (29) for the discharge of the capacitor.

Figure (5) appears less intuitive than Figure (3) because we have drawn a current flowing into the capacitor, while we know that the current actually flows out. But the fact that we analyzed the circuit in Figure (5), assuming the wrong direction for the current, does not affect the resulting differential equation for the circuit. When using Kirchoff's laws to analyze a circuit, you do not have to know the correct direction for the currents ahead of time. If you make the wrong guess, the resulting equations will fix things up by giving you a minus sign.

While Figure (5) is less intuitive than Figure (3), it is much more straightforward to stick with all positive quantities and always label our circuit element voltages and currents as shown in Figure (4). With more complex circuits it is the only way to maintain sanity and get the right differential equation.

**Figure 5** *The RC circuit for a positive increasing current i.* 



### Solving the RC Circuit Equation

Solving the differential equation

$$\frac{\mathrm{dQ}}{\mathrm{dt}} + \frac{\mathrm{Q}}{\mathrm{RC}} = 0 \qquad (29) \text{ repeated}$$

for the capacitor discharge was quite straightforward. We first looked at the circuit experimentally and saw that the voltage Q/C appeared to decay exponentially as shown in Figure (27-44c) from the Physics text, reproduced here. This suggested that we try, as a guess, a solution of the form

$$Q = Q_0 e^{-\alpha t} \tag{31}$$

$$\frac{\mathrm{d}Q}{\mathrm{d}t} = -\alpha Q_0 \mathrm{e}^{-\alpha t} \tag{32}$$

Plugging our guess into Equation (30) gives

$$-\alpha Q_0 e^{-\alpha t} + \frac{Q_0 e^{-\alpha t}}{RC} \neq 0$$

The constants  $Q_0$  and the functions  $e^{-\alpha t}$  cancel, and we are left with

$$-\alpha + \frac{1}{RC} \neq 0$$

We can satisfy the differential equation if  $\alpha$  has the value

$$\alpha = \frac{1}{\text{RC}}$$
(33)

The formula for Q becomes

$$Q = Q_0 e^{-t/RC}$$
(34)

We see the time constant for the decay of the charge Q is T = RC. I.e., when t gets up to T = RC, the value of the charge has decreased to  $e^{-1} = 1/e$  of its original value.



### Figure 27-44c

Discharge of our aluminum plate capacitor (separation 2mm) through a  $10K\Omega$  resistor. The inset is the experimental data and the solid curve is drawn from that data.

### The LC Circuit

We will construct an LC circuit from Figure (4) by taking out the resistor and connecting the back side as shown in Figure (6). Setting the sum of the voltage rises around this circuit equal to zero gives

$$\mathbf{V}_{\mathrm{L}} + \mathbf{V}_{\mathrm{C}} = \mathbf{0} \tag{35}$$

$$L\frac{di}{dt} + \frac{Q}{C} = 0 ; \quad i = +\frac{dQ}{dt}$$
(36)

Writing  $di/dt = d^2\,Q/dt^2\,$  , and dividing through by L gives

$$\frac{\mathrm{d}^2 \mathrm{Q}}{\mathrm{d}t^2} + \frac{\mathrm{Q}}{\mathrm{LC}} = 0 \tag{37}$$

Now suppose we naïvely try the same exponential decay solution we had for the RC circuit

$$Q = Q_0 e^{-\alpha t}; \quad \frac{dQ}{dt} = -\alpha Q_0 e^{-\alpha t}$$
 (guess)

$$\frac{d^2Q}{dt^2} = -\alpha Q_0 \left(-\alpha e^{-\alpha t}\right) = \alpha^2 Q_0 e^{-\alpha t} \qquad (38)$$

Plugging our guess (38) into the differential Equation (37) gives

$$\alpha^2 Q_0 e^{-\alpha t} + \frac{Q_0 e^{-\alpha t}}{LC} \stackrel{?}{=} 0$$

Again the  $Q_0$ 's and  $e^{-\alpha t}$ 's cancel and we are left with

$$\alpha^2 + \frac{1}{LC} \stackrel{\text{\tiny $2$}}{=} 0 \tag{39}$$

The differential equation will be solved if we can set

$$\alpha^2 = -\frac{1}{LC} \tag{40}$$

$$\alpha = -\frac{1}{\sqrt{LC}}\sqrt{-1}$$
 (41)





When we tried this in the Physics text, we noted that  $\alpha$  comes out imaginary. We also noted that the LC circuit oscillated rather than decayed. Thus we concluded that we had guessed the wrong function, and tried a sine wave

$$Q = Q_0 \sin \omega_0 t \tag{42}$$

instead. When you plug the guess (42) into the differential Equation (37) you end up with

$$\omega_0^2 = \frac{1}{\sqrt{LC}} ; \quad \omega_0 = \frac{1}{\sqrt{LC}}$$
(43)

which avoids imaginary numbers and gives a result in agreement with experiment. The quantity  $\omega_0 = 1/\sqrt{LC}$  is the resonant frequency of the LC circuit. (If you do not remember plugging the guess  $Q = Q_0 \sin(\omega_0 t)$  into the LC differential equation, do so now.)

Knowing more about handling imaginary numbers, let us see what happens if we take our guess  $Q = e^{-\alpha t}$ seriously for the LC circuit. We still have to satisfy Equation (40),

$$\alpha^2 = -1/LC$$
  
Writing  $1/LC = \omega_0^2$ , we get

$$\alpha^2 = -\omega_0^2 \tag{44}$$

which has two solutions, namely

$$\alpha = i\omega_0 \tag{45a}$$

$$\alpha = -i\omega_0 \tag{45b}$$

You can see this by noting that both  $i^2 = -1$  and  $(-i)^2 = -1$ . Thus the possible solutions for Q are

$$\mathbf{Q}_1 = \mathbf{Q}_0 \mathbf{e}^{i\,\omega_0 \mathbf{t}} \tag{46a}$$

$$Q_2 = Q_0 e^{-i\omega_0 t} \tag{46b}$$

While Equations (46a, b) are both mathematical solutions to the differential equation for the LC circuit, both are complex functions. But the amount of charge Q in the capacitor must be described by a real number. No imaginary charge resides there.

However in Equations (24) and (25) we saw that we could construct the real functions  $\cos\theta$  and  $\sin\theta$  from the complex functions  $e^{i\theta}$  and  $e^{-i\theta}$ . Replacing  $\theta$  by  $\omega_0 t$ , we have

$$\cos \omega_0 t = \frac{1}{2} \left( e^{i \omega_0 t} + e^{-i \omega_0 t} \right)$$
(47a)

$$\sin\omega_0 t = \frac{i}{2} \left( e^{-i\omega_0 t} - e^{i\omega_0 t} \right)$$
(47b)

One of the features of differential equations like the one for the LC circuit<sup>\*</sup> is that if the equation has more than one solution, any combination of the solutions is also a solution of the equation. In our case the two solutions are  $Q_1 = Q_0 e^{i\omega_0 t}$  and  $Q_2 = Q_0 e^{-i\omega_0 t}$ . Thus the combination

$$Q = aQ_1 + bQ_2 \qquad a,b,constants \qquad (48)$$

must also be a solution, as you will check for yourself in Exercise 2. Choosing the constants a = 1/2, b = +1/2 gives

$$Q = Q_0 \cos \omega_0 t \tag{49a}$$

and choosing a = -i/2, b = i/2 gives

$$Q = Q_0 \sin \omega_0 t \tag{49b}$$

These are both real functions which can describe the electric charge in the capacitor.

Thus we see that for both the RC and the LC circuit, we can use the same trial function  $Q = Q_0 e^{-\alpha t}$ . For the RC circuit,  $\alpha$  was a real number, which gave us the exponential decay  $Q = Q_0 e^{-t/RC}$ . For the LC circuit,  $\alpha$  turned out to be imaginary which gave us real oscillating solutions like  $Q = Q_0 \cos \omega_0 t$ . By using complex numbers, we are able to handle both the RC and the LC circuits with the same trial function. Whether  $\alpha$  turns out to be real or imaginary tells us whether the circuit decays or oscillates.

<sup>\*</sup>This is an example of what is called a *homogenous differential equation*. We will have more to say about them shortly.

In the next section we will consider the RLC circuit, which is an LC circuit with resistance included. Experimentally we saw that such a circuit could have a decaying oscillation. When we plug the guess  $Q = Q_0 e^{-\alpha t}$  into the equation for the RLC circuit,  $\alpha$  will turn out in some cases to be complex, i.e., have both a real and an imaginary part. The imaginary part will describe the oscillation of the circuit while the real part will tell us how the oscillation decays. But before we get to the RLC circuit, we need to discuss a simpler way to get real solutions from complex solutions of differential equations. Before that, do Exercise 2 to see that  $Q = aQ_1 + bQ_2$ is a solution of our LC equation.

### Exercise 2

The differential equation for an LC circuit is

 $\frac{d^2Q}{dt^2} + \frac{Q}{LC} = 0$  (37) repeated

This is called a *homogenous differential equation* because it contains only terms involving Q or its derivatives. An example of a *non homogeneous differential equation* will be

$$\frac{d^2Q}{dt^2} + \frac{Q}{LC} = a\sin\omega_1 t$$
(50)

This will represent an LC circuit that is being forced to oscillate at some frequency  $\omega_1$ . The appearance of the term (a sin  $\omega_1 t$ ) with no factor of Q makes this a *non homogeneous* equation. We will discuss this equation shortly, to show what effect the non homogeneous term has. For now we will limit our discussion to homogeneous equations.

You have seen that  $Q_1 = ae^{i\omega_0 t}$  and  $Q_2 = be^{-i\omega_0 t}$  are both solutions to Equation (37) when a and b are constants and  $\omega_0 = 1/\sqrt{LC}$ . Now explicitly plug in

$$Q = ae^{i\omega_0 t} + be^{-i\omega_0 t}$$
(51)

into Equation (37), and show that this is a solution for any constant values of a and b. This demonstrates that any linear combination of  $e^{i\omega_0 t}$  and  $e^{-i\omega_0 t}$  is also a solution.

### A FASTER WAY TO FIND REAL SOLUTIONS

When we got the complex solutions  $e^{+i\omega_0 t}$  and  $e^{-i\omega_0 t}$  for the LC circuit differential equation, we were careful to construct real combinations of these complex solutions. You might think that it was lucky that we just happened to know that the combination  $1/2(e^{i\omega_0 t} - e^{-i\omega_0 t})$  was the real function  $\cos\omega_0 t$ . You might be concerned that for some other differential equations you would not be so lucky.

Don't worry. If you find a complex solution for a homogeneous differential equation, you can simply take the real part of the complex solution and throw away the imaginary part. This works because both the real part and the imaginary part must separately be solutions of the differential equation. (You could also keep the imaginary part without the (i) and throw away the real part.)

To see why both the real and imaginary parts are solutions, let us write the complex solution for Q in the form

$$Q = Q_{real} + iQ_{imaginary}$$
(52)

where both  $Q_{real}$  and  $Q_{imaginary}$  are real functions. Plugging Equation (52) into the LC differential equation gives

$$\frac{d^2Q}{dt^2} + \frac{Q}{LC} = \left[\frac{d^2Q_{real}}{dt^2} + \frac{Q_{real}}{LC}\right] + i\left[\frac{d^2Q_{imaginary}}{dt^2} + \frac{Q_{imaginary}}{LC}\right]$$
$$= 0$$
(53)

Since both  $Q_{real}$  and  $Q_{imaginary}$  are real functions, their derivatives must also be real functions, and the quantities inside both square brackets in Equation (53) must be real. As a result the first square bracket is purely real, and the second square bracket with its factor of (*i*) must be purely imaginary. The only way you can add purely real and purely imaginary functions together to get zero is for both functions to be separately equal to zero.

That is, we must have

$$\frac{d^2 Q_{real}}{dt^2} + \frac{Q_{real}}{LC} = 0$$
 (54a)

$$\frac{d^2 Q_{imaginary}}{dt^2} + \frac{Q_{imaginary}}{LC} = 0$$
 (54b)

Equations (54) tell us that both functions  $Q_{real}$  and  $Q_{imaginary}$  must be solutions of the LC differential equation. If we want a real solution, we can use either  $Q_{real}$ ,  $Q_{imaginary}$  or any linear combination of the two. A similar argument applies to the solution of any homogeneous differential equation.

As an explicit example for our LC equation, suppose we had come up with the solution

$$Q = Q_0 e^{i\omega_0 t}$$
(55)

and had not noticed that  $Q = Q_0 e^{-i\omega_0 t}$  was also a solution. Instead of hunting for another complex solution and then trying to find real combinations, we could just break  $e^{i\omega_0 t}$  into its real and imaginary parts using  $e^{i\theta} = \cos \theta + i \sin \theta$  to get

$$Q = Q_0 e^{i\omega_0 t} = Q_0 \cos\omega_0 t + iQ_0 \sin\omega_0 t \quad (55a)$$

Then we immediately know that the real functions  $Q_0 \cos \omega_0 t$  and  $Q_0 \sin \omega_0 t$  are solutions of the LC differential equation. We can use either one or some linear combinations of the two. (Using a linear combination is equivalent to using an arbitrary phase angle, like  $Q = Q_0 \sin(\omega_0 t + \phi)$ . See the Physics text, pages 15-17 or 16-31.)



### THE RLC CIRCUIT

Adding a resistor to an LC circuit gives us the RLC circuit shown in Figure (7). If the resistance R is not too large, we get a decaying oscillation like that shown in Figure (31-A9) taken from the Physics text.

The equation for the RLC circuit is obtained by setting to zero the sum of the voltage rises around the circuit, giving

$$\mathbf{V}_{\mathbf{R}} + \mathbf{V}_{\mathbf{L}} + \mathbf{V}_{\mathbf{C}} = \mathbf{0} \tag{56}$$

$$iR + L\frac{di}{dt} + \frac{Q}{C} = 0$$
(57)

Setting

$$i = \frac{dQ}{dt}$$
;  $\frac{di}{dt} = \frac{d^2Q}{dt^2}$  (58)

and dividing through by L gives

$$\frac{d^2Q}{dt^2} + \frac{R}{L}\frac{dQ}{dt} + \frac{Q}{LC} = 0 \qquad the LRC \\ equation \qquad (59)$$

As a trial function, suggested by the decaying oscillation of Figure (31-A9), we could try the solution

$$Q = Q_0 e^{-\alpha t} \cos \omega t \qquad guess \qquad (60)$$



**Figure 31-A9 – Ringing like a bell** We hit the RLC circuit with a square wave and the circuit responded like a bell struck by a hammer. We are looking at the voltage across the capacitor.

If you plug the guess (60) into Equation (59), you get many terms involving both sin  $\omega t$  and  $\cos \omega t$ .

$$Q = Q_0 e^{-\alpha t} \cos \omega t \qquad guess \quad (60) \text{ repeated}$$
$$\frac{d^2 Q}{dt^2} + \frac{R}{L} \frac{dQ}{dt} + \frac{Q}{LC} = 0 \qquad (59) \text{ repeated}$$

To see where the terms come from, consider

$$\frac{\mathrm{d}Q}{\mathrm{d}t} = Q_0(-\alpha)e^{-\alpha t}\cos\omega t + Q_0e^{-\alpha t}(-\sin\omega t)$$
(61)

where we had to differentiate the two terms  $e^{-\alpha t}$  and  $\cos \omega t$  separately. Differentiating again we get four terms for  $d^2Q/dt^2$ , two with a  $\cos \omega t$  and two with a  $\sin \omega t$ . When we plug this all back into Equation (59), we end up with seven terms, four with  $\cos \omega t$  and three with  $\sin \omega t$ . In order for all this to be equal to zero, you have to separately set the  $\sin \omega t$  and the  $\cos \omega t$  terms equal to zero. This leads to two equations, from which you can determine both the constants  $\alpha$  and  $\omega$ . If you are careful, your chances of getting the answer without making a mistake may be as high as 50%. In other words this is the hard way to solve the problem.

### **Exercise 3**

Try finding the coefficients  $\alpha$  and  $\omega$  by using Equation (60) as a trial solution for Equation (59). Then check your answer with the one we get in the next section.



**Figure 31-A9 (repeated)** We are looking at the voltage across the capacitor in an RLC circuit.

### The Easy Way

Working with separate sines and cosines is the difficult way to handle the RLC circuit. Using complex variables which provide a unified treatment of both decay and oscillation is the easy way.

For a trial solution, let us use

$$Q = Q_0 e^{-at} ; \quad \frac{dQ}{dt} = -aQ_0 e^{-at}$$
$$\frac{d^2Q}{dt^2} = a^2 Q_0 e^{-at}$$
(62)

It looks much easier already. Substituting this trial solution into the LCR differential equation gives

$$\frac{d^2Q}{dt^2} + \frac{R}{L}\frac{dQ}{dt} + \frac{Q}{LC} = 0$$
 (59) repeated

$$a^{2}Q_{0}e^{-at} - \frac{aR}{L}Q_{0}e^{-at} + \frac{Q_{0}e^{-at}}{LC} \stackrel{2}{\neq} 0$$
 (63)

The function  $e^{-at}$  and constant  $Q_0$  cancel and we are left with

$$a^2 - \frac{aR}{L} + \frac{1}{LC} \stackrel{2}{\neq} 0 \tag{64}$$

This is a standard quadratic equation of the form

$$x^2 + bx + c = 0 (65)$$

whose solution is

$$x = \frac{-b \pm \sqrt{b^2 - 4c}}{2}$$
(66)

For our case, -b = R/L, c = 1/LC, thus (a) is given by

$$a = \frac{1}{2} \left( \frac{R}{L} \pm \sqrt{\frac{R^2}{L^2} - \frac{4}{LC}} \right)$$
$$= \frac{R}{2L} \pm \sqrt{\frac{R^2}{4L^2} - \frac{1}{LC}}$$
(67a)

$$= \frac{\mathbf{R}}{2\mathbf{L}} \pm \sqrt{(-1)\left(\frac{1}{\mathbf{LC}} - \frac{\mathbf{R}^2}{4\mathbf{L}^2}\right)}$$
(67b)

Setting  $1/LC = \omega_0^2$ , where  $\omega_0$  is the resonant frequency of the undamped (R = 0) circuit, and taking  $\sqrt{-1}$  outside the square root as a factor of (*i*) gives

$$a = \frac{R}{2L} \pm i \sqrt{\left(\omega_0^2 - \frac{R^2}{4L^2}\right)}$$
(68)

We now introduce the notation

$$\alpha = \frac{R}{2L}$$

$$\omega = \sqrt{\omega_0^2 - \frac{R^2}{4L^2}} = \sqrt{\omega_0^2 - \alpha^2}$$
(69)

So that

$$a = \alpha \pm i\omega \tag{70}$$

and our trial solution  $Q = Q_0 e^{-at}$  becomes

$$Q = Q_0 e^{-(\alpha \pm i\omega)t}$$

$$Q = Q_0 e^{-\alpha t} e^{\pm i\omega t}$$
*solution for the RLC circuit*
(71)

where

$$\alpha = \frac{R}{2L}$$
;  $\omega = \sqrt{\omega_0^2 - \alpha^2}$  (69) repeated

For the case where  $\omega_0$  is bigger than  $\alpha$ ,  $\omega = \sqrt{\omega_0^2 - \alpha^2}$  is a real number, the real part of  $e^{i\omega t}$  is cos  $\omega t$  and we get the real solution

$$Q_1 = Q_0 e^{-\alpha t} \cos \omega t \tag{72}$$

The imaginary part of  $e^{i\omega t}$  is proportional to  $\sin\omega t$ , which gives us the other real solution

$$Q_2 = Q_0 e^{-\alpha t} \sin \omega t \tag{73}$$

As in the case of the LC circuit, the sine and cosine waves can be combined as a sine wave with an arbitrary phase angle  $\phi$  to give the general solution

$$Q = Q_0 e^{-\alpha t} \sin(\omega t + \phi) \qquad damped oscillation of an RLC circuit (74)$$

Equation (74) represents a damped oscillation of frequency  $\omega = \sqrt{\omega_0^2 - \alpha^2}$  and a damping time constant T given by

$$T = \frac{1}{\alpha} = \frac{2L}{R} \qquad \begin{array}{c} damping time \\ constant \end{array}$$
(75)

Imagine that we start with an RLC circuit that initially has negligible resistance, and that we gradually increase the resistance. When R = 0, then  $\alpha = 0$  and the oscillation frequency is  $\omega = \sqrt{\omega_0^2} = \omega_0$ , where  $\omega_0$  is the undamped frequency.

As R and  $\alpha = R/2L$  are increased, the oscillation frequency  $\omega = \sqrt{\omega_0^2 - \alpha^2}$  decreases until we reach  $\alpha = \omega_0$ . At that point,  $\omega = \sqrt{\omega_0^2 - \alpha^2} = 0$ , oscillation ceases, and we have what is called *critical damping*. The time constant for decay at critical damping is just the length of time it takes the undamped circuit to go through one radian of oscillation, or  $1/2\pi$  of a complete cycle. You can see that result from dimensions. We have

$$\frac{1}{\omega_0 \frac{\text{radians}}{\text{second}}} = \frac{1}{\omega_0} \frac{\text{seconds}}{\text{radian}}$$
(76)

and at critical damping, where  $\alpha = \omega_0$ ,

$$T = \frac{1}{\alpha} = \frac{1}{\omega_0} \frac{\text{seconds}}{\text{radian}}$$
(77)

At critical damping, there is only one unique solution for the RLC circuit. As we increase the resistance beyond critical damping, when  $\alpha = R/2L$  becomes larger than  $\omega_0$ , the solution becomes *overdamped*. For  $\alpha > \omega_0$ , it is easiest to go back to writing the solution in the form

$$Q = Q_0 e^{-at} \qquad (from Eq.62)$$

$$a = \frac{R}{2L} \pm \sqrt{\frac{R^2}{4L^2} - \frac{1}{LC}} = \alpha \pm \sqrt{\alpha^2 - \omega_0^2}$$
(from Eq.67a)

and we see that we now have two exponential decay solutions

$$Q_1 = Q_0 e^{-\left(\alpha + \sqrt{\alpha^2 - \omega_0^2}\right)t}$$
(78a)

$$Q_2 = Q_0 e^{-\left(\alpha - \sqrt{\alpha^2 - \omega_0^2}\right)t}$$
(78b)

If we increase the resistance so much that  $\omega_0^2$  is completely negligible compared to  $\alpha^2$ , then the two solutions become

$$Q_1 \rightarrow Q_0 e^{-2\alpha t} \qquad \alpha^2 >> \omega_0^2 \qquad (79a)$$

$$Q_2 \rightarrow Q_0 e^0 = Q_0 \tag{79b}$$

In this limit we easily see that the solution  $Q_1$  damps more rapidly than  $Q_2$ . For the  $Q_2$  solution, we have increased the resistance so much that no charge leaves the capacitor and the charge remains at  $Q_0$ . We can get a better insight into the solution  $Q_2$  by assuming that  $\omega_0^2$  is small but not quite zero. In this case we can write  $Q_2$  as

$$Q_2 = Q_0 e^{-\alpha \left(1 - \sqrt{1 - \omega_0^2 / \alpha^2}\right)t}$$
(80)

Since  $\omega_0^2/\alpha^2 \ll 1$ , we can use the approximation formula

$$\sqrt{1-x} \approx 1 - \frac{x}{2} \qquad x \ll 1$$

We get, for  $x = \omega_0^2/\alpha^2$ ,

$$\alpha \left[ 1 - \sqrt{1 - \frac{\omega_0^2}{\alpha^2}} \right] \approx \alpha \left[ 1 - \left( 1 - \frac{\omega_0^2}{2\alpha^2} \right) \right]$$
$$= \alpha \left[ \frac{\omega_0^2}{2\alpha^2} \right] = \frac{1}{2} \omega_0^2 \frac{1}{\alpha}$$

With  $\omega_0^2 = 1/LC$  and  $\alpha = R/2L$ , we get

$$= \frac{1}{2} \times \frac{1}{LC} \times \frac{2L}{R}$$
$$= \frac{1}{RC}$$
(81)

)

Thus for  $\alpha^2 >> \omega_0^2$  we have

$$Q_2 = Q_0 e^{-t/RC}$$
  $\alpha^2 >> \omega_0^2$  (82)

This is just the solution for the decay of an RC circuit with a time constant T = RC.

The condition  $\alpha^2 \gg \omega_0^2$  can be written as

$$\frac{\mathbf{R}^2}{4\mathbf{L}^2} \gg \frac{1}{\mathbf{LC}} \quad \text{or} \quad \frac{\mathbf{R}^2 \mathbf{C}}{4\mathbf{L}} \gg 1 \tag{83}$$

We can meet this condition for finite values of R and C by making L small enough.

### **Exercise 4**

To make our study of the RLC circuit more concrete, suppose that in the circuit you use a 0.10 microfarad capacitor and one millihenry inductor, so that

$$L = 10^{-3} \text{ hy}$$
$$C = 10^{-5} \text{ farads}$$

(a) What is the resonant frequency  $\omega_0$  radians/second and  $f_0$  cycles/second, when R = 0?

(b) What is the length of time it takes the R = 0 circuit to go through one radian of its oscillation?

(c) What value of resistance R<sub>C</sub> should you use for critical damping?

(d) What is the time constant for the decay at critical damping?

(e) Suppose you raise R from its critical value R<sub>C</sub> up to  $\sqrt{2}$ R<sub>C</sub>. What are the time constants T<sub>1</sub> and T<sub>2</sub> for the decay of the solutions Q<sub>1</sub> and Q<sub>2</sub> respectively? (Partial answer: Q<sub>2</sub> takes twice as long as Q<sub>1</sub> to decay when R =  $\sqrt{2}$ R<sub>C</sub>.)

### IMPEDANCE

Circuits commonly encountered are *AC circuits* where the current has a sinusoidal form

$$i = i_0 \sin \omega t$$
 (84)

For standard American households, the household current has a frequency of 60 cycles/second, or  $\omega = 2\pi \times 60$  radians/second. In much of the rest of the world the standard household frequency is 50 cycles per second. World War II aircraft used a standard frequency of 400 cycles per second which resulted in smaller and lighter transformers.

The concept of impedance, which involves complex variables, provides an easy way to handle the voltages across R, L, and C circuit elements in an AC circuit. To demonstrate the advantage of the complex variable approach, we will first analyze these voltages using our standard real variables, and then see how much the calculations are simplified by complex variables.

Suppose we have three circuit elements, an R, L, and C, connected in series as shown in Figure (8), and run an AC current through them. In the diagram we show the formula for the voltage across each circuit element. What we wish to calculate is the total voltage V across all three elements.

$$V = ?$$

$$V_{R} = iR = i_{0}Rsin\omega t (85a)$$

$$V = ?$$

$$V_{L} = L\frac{di}{dt} = i_{0}L\omega cos\omega t (85b)$$

$$C = \frac{Q}{C} = \frac{1}{C}\int idt$$

$$= \frac{-i_{0}}{C\omega} cos\omega t (85c)$$

AC voltages in the R, L, and C circuit elements.

The individual voltages were calculated noting that

$$\frac{d\sin\omega t}{dt} = \omega\cos\omega t$$
$$\int \sin\omega t \, dt = -\frac{1}{\omega}\cos\omega t.$$

The voltage V across all three elements is just the sum of the individual voltages

$$V = i_0 \left( R \sin \omega t + L \omega \cos \omega t - \frac{1}{C \omega} \cos \omega t \right)$$
$$= i_0 \left[ R \sin \omega t + \left( L \omega - \frac{1}{\omega C} \right) \cos \omega t \right]$$
$$V = i_0 \left[ A \sin \omega t + B \cos \omega t \right]$$
(86)

where

A = R; B = 
$$\left(L\omega - \frac{1}{\omega C}\right)$$
 (87)

We want to express the term  $[A \sin \omega t + B \cos \omega t]$  as a single sine wave with an amplitude which we will call Z<sub>0</sub>, and a phase angle  $\phi$ 

$$[A\sin\omega t + B\cos\omega t] = Z_0\sin(\omega t + \phi) \qquad (88)$$

To do this we use the trigonometric identity

$$sin(a + b) = cos b sin a + sin b cos a$$

to write

$$\sin(\omega t + \phi) = \cos\phi \sin\omega t + \sin\phi \cos\omega t \quad (89)$$

Multiplying through by  $Z_0$  gives

$$Z_0 \sin(\omega t + \phi) = (Z_0 \cos \phi) \sin \omega t + (Z_0 \sin \phi) \cos \omega t$$

$$= A\sin\omega t + B\cos\omega t \qquad (90)$$

where

$$A = Z_0 \cos \phi; \qquad B = Z_0 \sin \phi \qquad (91)$$

$$\frac{B}{A} = \frac{\sin\phi}{\cos\phi} = \tan\phi \tag{92}$$

$$A^{2} + B^{2} = Z_{0}^{2}(\cos^{2}\phi + \sin^{2}\phi) = Z_{0}^{2}$$
 (93)

Applying Equations (91), (92), and (93) to our formula

$$V = i_0 [A \sin \omega t + B \cos \omega t]$$
 (86) repeated gives

$$V = i_0 Z_0 \sin(\omega t + \phi)$$
 (94)

where from Equations (92) and (87)

$$\tan\phi = \frac{B}{A} = \frac{L\omega - 1/\omega C}{R}$$
(95)

and from Equation (93)

$$Z_0^2 = R^2 + \left(L\omega - \frac{1}{\omega C}\right)^2$$
(96)

After a fair amount of calculation, we see that the voltage across all three circuit elements is still proportional to  $\sin \omega t$ . Its amplitude  $Z_0$  is given by Equation (96) and there is a phase shift by an angle  $\phi$  that is given by Equation (95).



AC voltages in the R, L, and C circuit elements, using complex notation.

Now let us see how much more quickly we can arrive at the amplitude  $Z_0$  and phase shift  $\phi$  using the complex variables shown in Figure (9).

In Figure (9) we have a current i given by the formula

$$\mathbf{i} = \mathbf{i}_0 \mathbf{e}^{i\,\boldsymbol{\omega}\mathbf{t}} \tag{97}$$

and the resulting voltage across the three circuit elements is the sum of the individual voltages which can easily be written in the form

$$\mathbf{V} = \mathbf{i}_0 \left[ \mathbf{R} + i \left( \mathbf{L} \boldsymbol{\omega} - \frac{1}{\boldsymbol{\omega} \mathbf{C}} \right) \right] \mathbf{e}^{i \, \boldsymbol{\omega} \mathbf{t}} \tag{98}$$

The quantity in square brackets is the complex number  $R + i(L\omega - 1/\omega C)$  graphed in Figure (10). It can be represented by an arrow whose length is  $Z_0$  given by the Pythagorean theorem as

$$Z_0^2 = R^2 + \left(L\omega - \frac{1}{\omega C}\right)^2$$
(99)

and is oriented at an angle  $\phi$  whose tangent is

$$\tan \phi = \frac{L\omega - 1/\omega C}{R} \tag{100}$$

Notice that the formulas for  $Z_0$  and  $\tan \phi$  are the same as in Equations (96) and (95), which we got after so much more work.



**Figure 10** Graph of the complex number  $R + i(L\omega - 1/\omega C)$ .

From our earliest work with complex variables we saw that the complex number

z = x + iy (4) repeated

could be written as the exponential

 $z = re^{i\theta}$  (26) repeated

where z is graphed in Figure (2a) repeated here.

Thus the complex number  $R + i(L\omega - 1/\omega C)$ , graphed in Figure (10) can also be written in the exponential form

$$\mathbf{R} + i \left( \mathbf{L} \boldsymbol{\omega} - \frac{1}{\boldsymbol{\omega} \mathbf{C}} \right) = \mathbf{Z}_0 \mathbf{e}^{i\theta} \tag{101}$$

where  $Z_0$  is the distance from the origin and  $\phi$  the angle above the real axis.

Using Equation (100) for the square brackets in Equation (98) for the voltage V gives

$$V = i_0 \left[ R + i \left( L \omega - \frac{1}{\omega C} \right) \right] e^{i \omega t}$$
  
=  $i_0 \left[ Z_0 e^{i \phi} \right] e^{i \omega t}$   
$$V = i_0 Z_0 e^{i (\omega t + \phi)}$$
(102)

$$Z_0^2 = R^2 + \left(L\omega - \frac{1}{\omega C}\right)^2$$
 (99) repeated

$$\tan \phi = \frac{L\omega - 1/\omega C}{R}$$
(100) repeated



**Figure 2a (repeated)** Plot of the complex number (4 + 3i), showing the angle  $\phi$ .

Equation (102) is our complex formula for the voltage across the three circuit elements.

To find the real voltage, we simply take the real (or imaginary) part of the complex voltage. Choosing the imaginary part (without the i) to get a sine wave, we get

$$V = i_0 Z_0 \sin(\omega t + \phi)$$
 (103)

which is the same answer, Equation (94), that we got from the real analysis.

The main advantage of the complex analysis is that all the voltages had the same factor  $e^{i\omega t}$ , so that we could simply add the voltages without using the fairly messy trigonometric identities. Also note that the main result of all the work of the real analysis was to calculate the amplitude  $Z_0$  and the phase angle  $\phi$ . We got  $Z_0$  and  $\phi$  immediately in the complex analysis, as soon as we graphed the complex coefficient of  $e^{i\omega t}$  in Figure (10).

### **Impedance Formulas**

The concept of a *complex impedance* which we will now introduce, allows you to determine the amplitude  $Z_0$  and phase angle  $\phi$  by inspection, without doing hardly any calculation at all.

In Figure (11), we have redrawn our three circuit elements, introduced a complex current  $i = i_0 e^{i\omega t}$ , and expressed voltage in terms of i and the complex impedances  $Z_R$ ,  $Z_L$ ,  $Z_C$  defined by

$$Z_R \equiv R \tag{104a}$$

$$Z_{\rm L} \equiv i\omega L \tag{104b}$$

$$Z_{\rm C} \equiv -\frac{i}{\omega \rm C} \tag{104c}$$

In terms of these Z's, the voltages are

$$V_{R} = iZ_{R}$$

$$V_{L} = iZ_{L}$$

$$V_{C} = iZ_{C}$$
(105)

The sum of the three voltages V becomes

$$V = V_R + V_L + V_C$$
$$= i(Z_R + Z_L + Z_C)$$
(106)



### Figure 11

The voltages  $V_R$ ,  $V_L$ , and  $V_C$  expressed in terms of impedances Z. If we define the total impedance Z of the three circuit elements connected in series by the equation

$$Z = Z_{R} + Z_{L} + Z_{C}$$
(107)

then our formula for the complex voltage is

$$V = iZ$$
(108)

Comparing this with Ohm's law for a single resistor

$$V_R = iR$$
 Ohm's law (Physics 27-1)

we see that we can think of Equation (108) as simply a complex form of Ohm's law.

When we graph the complex impedance Z we can immediately read off the amplitude  $Z_0$  and phase angle  $\phi$ , as shown in Figure (12). We have

$$Z = R + i \left( L\omega - \frac{1}{\omega C} \right) = Z_0 e^{i\phi} \quad \begin{array}{c} complex \\ impedance \end{array}$$
(109)

where

$$Z_0^2 = R^2 + \left(L\omega - \frac{1}{\omega C}\right)^2$$
 magnitude of (110a) impedance

$$\tan \phi = \frac{L\omega - 1/\omega C}{R} \qquad phase of \\ impedance \qquad (110b)$$

In Equation (109), we introduced the exponential form  $Z_0 e^{i\phi}$  for the complex variable Z.





The complex impedance can be pictured as an arrow of length  $Z_0 = \sqrt{Z_{real}^2 + Z_{imag}^2}$  oriented at an angle  $\phi$ .

### The Driven RLC Circuit

Our first demonstration in the physics course was the driven RLC circuit, which could be used to measure the speed of light without looking at light. (This was a crucial point in our discussion of special relativity.) In Chapter 31 we calculated the resonant frequency of an LC circuit and wrote down some formulas for the driven RLC circuit. But we did not derive the formulas because the work is messy when we have to use real functions. However with the complex analysis we have developed in this chapter, we get, almost by inspection, not only the formulas but considerable insight into the behavior of the circuit.

In the lecture demonstration, we drove the LRC circuit by wrapping a couple of turns of wire around the outside of the inductor and attaching the wire to an oscillator. The oscillating magnetic flux produced by these few turns induces a voltage  $V_{ind}$  in the coil and drives the circuit to oscillate.

The important thing is that we did not put the oscillator directly in the circuit, for the oscillator has its own internal resistance, capacitance and inductance that could completely alter the behavior of the circuit. The idea is to give the circuit a gentle voltage shove of the form

$$V_{\rm ind} = V_0 e^{i\omega t} \tag{111}$$

as indicated in Figure (13), and see how the circuit responds.



Figure 13 The driven RLC circuit. Photo is Figure (1-10) from the Physics text.



Setting the sum of the voltage rises to zero around the circuit in Figure (13) gives, (walking counter clockwise),

$$V_{\rm C} + V_{\rm L} + V_{\rm R} - V_{\rm ind} = 0$$
 (112a)

$$iZ_{R} + iZ_{L} + iZ_{C} = V_{0}e^{i\omega t}$$
(112b)

Solving for the current i in the circuit gives

$$i = \frac{V_0 e^{i\omega t}}{Z}$$
(113)

where  $Z = Z_R + Z_L + Z_C = Z_0 e^{i\phi}$  is the total impedance of the circuit.

Using the exponential form for Z in Equation (113) for the current i gives

$$\mathbf{i} = \frac{\mathbf{V}_0 \mathbf{e}^{i\,\boldsymbol{\omega}\mathbf{t}}}{\mathbf{Z}_0 \mathbf{e}^{i\boldsymbol{\phi}}}; \qquad \mathbf{i} = \left(\frac{\mathbf{V}_0}{\mathbf{Z}_0}\right) \mathbf{e}^{i\,(\boldsymbol{\omega}\mathbf{t}-\boldsymbol{\phi})} \qquad (114)$$

Equation (114) tells us that if we drive an RLC circuit with an induced voltage  $V_{ind} = V_0 e^{i\omega t}$  the circuit will respond with a current i that has an amplitude  $(V_0/Z_0)$  and a phase  $(-\phi)$  relative to the driving voltage. We get this result almost without doing any calculation. To get the same result using real functions sin  $\omega t$  and cos  $\omega t$  would have taken several pages of algebra and trigonometric identities.



**Figure 14** *Complex impedance for an RLC circuit.* 

Let us look at the physics contained in Equation (114).

$$i = \left(\frac{V_0}{Z_0}\right) e^{i(\omega t - \phi)}$$
(114) repeated

For very low frequencies, for sufficiently small  $\omega$ , the quantity  $1/\omega C$  is much larger than either  $L\omega$  or R, the impedance is essentially all capacitive as indicated in Figure (15). For this case,

$$Z_0 \approx \frac{1}{\omega C}; \quad \phi \approx -90^\circ = -\frac{\pi}{2}$$
 (115)

and the formula for the current in the circuit caused by the induced voltage  $V_{ind}$  is

$$i = V_0 \omega C e^{i(\omega t + \pi/2)}$$

$$\begin{array}{c} current \\ at low \\ frequencies \end{array}$$
(116a)

$$V_{\text{ind}} = V_0 e^{i\omega t} \qquad \begin{array}{c} complex \\ induced \\ voltage \end{array} (116b)$$

Taking the real part of Equations (116) gives us the real current for a real induced voltage

$$i = V_0 \omega C \cos (\omega t + \pi/2)$$

$$V_{ind} = V_0 \cos \omega t$$
(117)

From Equations (117), we see that at low frequencies, the phase of the current is  $\pi/2$  ahead of the induced voltage, and the amplitude goes to zero as  $\omega$  goes to zero.

The other extreme, at high frequencies where  $\omega L$  is much bigger than R or  $1/\omega C$ , we have

$$Z_0 \approx L\omega \tag{118}$$

$$\phi \approx +90^{\circ} \quad (\pi/2) \tag{119}$$

And we get

$$i = \frac{V_0}{L\omega} e^{i(\omega t - \pi/2)} \qquad \begin{array}{c} current\\ at high\\ frequencies \end{array} \qquad (120)$$

Taking the real part gives us the real current

$$i = \frac{V_0}{L\omega} \cos (\omega t - \pi/2)$$

$$V_{ind} = V_0 \cos \omega t$$
*large*  $\omega$  (121)

We see that at high frequencies the phase of the current is  $\pi/2$  behind of the induced voltage, and the amplitude goes to zero as  $\omega$  goes to infinity.





There is a special frequency, call it  $\omega_0$ , where the capacitive impedance  $Z_C = -1/\omega_0 C$  just cancels the inductive impedance  $Z_L = L\omega_0$ , leaving us with a pure resistive impedance  $Z_R = R$ , as shown in Figure (17).

This happens when

$$Z_{L} = -Z_{C}$$

$$\omega_{0}L = +\frac{1}{\omega_{0}C}$$
(122)

$$\omega_0^2 = \frac{1}{LC}$$
(123)

This special frequency is the resonant frequency  $\omega_0 = 1/\sqrt{LC}$  of the RLC circuit. We now see that the resonance occurs when the capacitive and inductive impedances cancel, leaving only the resistance to dampen the current in the circuit. Also note that at this frequency the phase angle  $\phi$  is zero, and the current i is given by



Figure 17

At resonance, the capacitive and inductive impedances cancel, and we are left with only the resistive impedance.

Taking the real part of Equation 24 gives

$$i = \frac{V_0}{R} \cos \omega_0 t$$

$$V_{ind} = V_0 \cos \omega_0 t$$

$$at_{resonance} (125)$$

We see that, at resonance, the current and the induced voltage are in phase with each other, and the only thing that limits the current is the actual resistance R in the circuit.

Comparing Equations (117, 121, and 125), we see that the phase of the current shifts by 180 degrees ( $\pi$ ) as we go from well below to well above the resonance. The smaller the value of R, the sharper the resonance, and the faster this phase shift occurs. The shape of the resonance curves, for three different values of R were shown in the Physics text, Figure (14-31) repeated here.



### Figure 14-31

Amplitude of the oscillation for various values of the resistance R. The peak occurs at  $\omega = \omega_0$  because the inductive and capacitive impedances cancel at the resonant frequency  $\omega_0$ .

### TRANSIENTS

While the above discussion of the driven RLC circuit describes what you most likely will see when you study the circuit in the lab, it is not the whole story. There are other solutions for the circuit, solutions which die out as time goes on, and thus are called *transient* solutions. To see where the transients come from, we need to go back to the differential equation for the driven circuit. We get the equation from Figure (18) which is simply Figure (13) with some labels changed. To make the circuit more nearly what we deal with in the lab, we are writing the induced voltage as a real function  $V_0 \cos \omega_d t$ , where we are now calling the driving frequency  $\omega_d$ .

### **Particular Solution**

Setting the sum of the voltages around the circuit equal to zero gives

$$V_R + V_L + V_C = V_{ind}$$
 (110) repeated

$$i\mathbf{R} + L\frac{di}{dt} + \frac{Q}{C} = V_0 \cos \omega_d t$$
 (126)

This time, let us express everything in terms of the current i rather than the charge Q, by differentiating Equation (126) with respect to time and using i = dQ/dt. We get, after dividing through by L

$$\frac{d^{2}i}{dt^{2}} + \frac{R}{L}\frac{di}{dt} + \frac{i}{LC} = \left(\frac{-V_{0}\omega_{d}}{L}\right)\sin\omega_{d}t \quad (127)$$

where we used  $d(\cos \omega_d t)/dt = -\omega_d \sin \omega_d t$ .



**Figure 18** *The driven RLC circuit again.* 

Equation (127) is an example of a *non-homogeneous differential equation*. It is non-homogeneous because of the driving term  $(-V_0\omega_d/L) \sin \omega_d t$  which does not have a factor of the variable (i) or a derivative of (i). This is called the *inhomogeneous term*.

In the previous section, we found that Equation (127) has the solution

$$i_{p} = \frac{V_{0}}{Z_{0}} e^{i(\omega t - \phi)} \qquad particular \\ solution \qquad (114) repeat$$

where

$$Z_0^2 = R^2 + (L\omega - \frac{1}{\omega C})^2 \qquad (99) \text{ repeated}$$

$$\tan \phi = \frac{L\omega - 1/\omega C}{R}$$
(100) repeated

The value of  $i_p$  from Equation (113) is called the *particular solution* of the differential equation (127).

### **Transient Solutions**

To see what the other solutions are, let us look at the *homogeneous differential equation* 

$$\frac{\mathrm{d}^2 \mathrm{i}}{\mathrm{d}t^2} + \frac{\mathrm{R}}{\mathrm{L}}\frac{\mathrm{d}\mathrm{i}}{\mathrm{d}t} + \frac{\mathrm{i}}{\mathrm{LC}} = 0 \tag{128}$$

which represents an RLC circuit with no driving term. I.e., it is Equation (127) without the *inhomogeneous* term.

As a review, let us see how quickly we can solve Equation (128). Using the trial solution

$$i = i_0 e^{-at}$$
;  $\frac{di}{dt} = -ae^{-at}$ ;  $\frac{d^2i}{dt^2} = a^2 e^{-at}$ 

gives

$$a^2 - \frac{R}{L}a + \frac{1}{LC} = 0$$

This is a quadratic equation in a, of the form  $a^2+ba+c = 0$  which has the solution

$$a = \frac{-b \pm \sqrt{b^2 - 4c}}{2} = \frac{-b}{2} \pm \sqrt{\frac{b^2}{4} - c}$$

With b = R/L and c = 1/LC we get  
a = 
$$\frac{R}{2L} \pm \sqrt{\frac{R^2}{4L^2} - \frac{1}{LC}}$$
  
=  $\frac{R}{2L} \pm i\sqrt{\frac{1}{LC} - \frac{R^2}{4L^2}}$ 

Thus the solution to Equation (128) is

$$i_{\rm T} = i_0 e^{-\alpha t} e^{\pm i\omega t} \qquad so called transient solution \qquad (129)$$

Where

$$\alpha = \frac{R}{2L} ; \quad \omega = \sqrt{\frac{1}{LC} - \frac{R^2}{4L^2}}$$

We can write  $\omega$  in the form

$$\omega^2 = \omega_0^2 - \alpha^2$$

where

$$\omega_0 = \frac{1}{\sqrt{LC}}$$

is the resonant frequency. Equation (129) is just Equation (71) expressed in terms of the current i rather than the charge Q. We are calling this a *transient solution*  $i_T$ . The reason for the name will become apparent shortly.

### **Combined Solutions**

Let us now go back to Equation (127) for the driven circuit, and write  $i_d$  for the constant ( $-V_0\omega_d/L$ ) in order to simplify the equation's appearence

$$\frac{d^2i}{dt^2} + \frac{R}{L}\frac{di}{dt} + \frac{i}{LC} = i_d \sin \omega_d t \qquad (127a)$$

Now try the solution

$$i_{\text{new}} = i_{\text{p}} + ai_{\text{T}} \tag{130}$$

where  $i_p$  is the particular solution (113),  $i_T$  is the transient solution of Equation (129), and (a) is an arbitrary constant. We know that

$$\frac{d^{2}(i_{p})}{dt^{2}} + \frac{R}{L}\frac{d(i_{p})}{dt} + \frac{(i_{p})}{LC} = i_{d}\sin\omega_{d}t \qquad (131)$$
non-homogeneous equation

$$\frac{d^{2}(ai_{T})}{dt^{2}} + \frac{R}{L}\frac{d(ai_{T})}{dt} + \frac{(ai_{T})}{LC} = 0$$
(132)  
homogeneous equation

Adding Equations (131) and (132) together gives

$$\frac{d^{2}(i_{p}+ai_{T})}{dt^{2}} + \frac{R}{L}\frac{d(i_{p}+ai_{T})}{dt} + \frac{(i_{p}+ai_{T})}{LC}$$
$$= i_{d}\sin\omega_{d}t$$
(133)

and we see that  $i_{new} = (i_p + ai_T)$  obeys the same equation as  $i_p$  alone. Thus  $i_{new}$  is a solution of the equation of the driven RLC circuit, for any value of the constant (a).

This result tells us that to the driven or particular solution  $i_p$ , we can add any amount of the homogeneous solution  $i_T$ , and we still have a solution for the driven RLC circuit.

The solutions  $i_T$  for the homogeneous equation are fundamentally different from the particular solution  $i_p$ . The driven solution

$$i_p = \frac{V_0}{Z_0} e^{i(\omega t - \phi)}$$
 (113) repeated

goes on at a constant amplitude  $V_0/Z_0$  for as long as the driving voltage is attached. The transient solution

$$i_{\rm T} = i_0 e^{-\alpha t} e^{i\omega t}$$
 (129) repeated

dies out exponentially with a time constant  $T = 1/\alpha$ . Because such solutions do not last, they are called *transient solutions*.

What you will observe in the lab is the following. When you first turn on or suddenly change the driving voltage  $V_0 \cos \omega_d t$ , you will see not only the particular solution  $i_p$ , but also some transients mixed in. If you wait for several time constants  $T = 1/\alpha$ , and keep the driving voltage amplitude  $V_0$  constant, the transients will die out and the pure driven solution will appear on your oscilloscope. If you want to see the transient solutions, you have to look within a time constant  $1/\alpha$  of the time you changed the driving voltage.

This finishes our discussion of the application of complex variables to the analysis of circuits. We now move on to the use of complex variables to describe wave motion.

### SOLUTIONS OF THE ONE DIMENSIONAL WAVE EQUATION

In Chapter 2 of the Calculus text we discussed the one dimensional wave equation applied to both waves on a rope and sound waves. Applied to waves on a rope, the equation was

$$\frac{\partial^2 \mathbf{y}(\mathbf{x}, \mathbf{t})}{\partial \mathbf{t}^2} = \mathbf{v}_{\text{wave}} \frac{\partial^2 \mathbf{y}}{\partial \mathbf{x}^2}$$
(134)

(Calculus 2-73)

where y(x,t) represented the height of the rope above its equilibrium position at some point x along the rope at some time t. (For a sound wave, replace y(x,t) by p(x,t) where p(x,t) is the change in pressure due to the sound wave at some point x and time t.)

(Recall that when we are working with more than one variable, like x and t, we use the notation  $\partial f(x,t)/\partial t$  to mean the derivative of f(x,t) with respect to t, holding x constant. This is called a *partial derivative* with respect to time).

We solved Equation (134) with a trial function of the form

$$y(x,t) = A \sin(kx - \omega t)$$
(135)

$$\frac{\partial^2 y}{\partial x^2} = -k^2 y \; ; \quad \frac{\partial^2 y}{\partial t^2} = -\omega^2 y \qquad (136)$$

to get

$$-\omega^{2}y = -v_{wave}^{2}k^{2}y$$

$$v_{wave}^{2} = \frac{\omega^{2}}{k^{2}}$$

$$v_{wave} = \frac{\omega}{k}$$
(137)

In the solution  $\sin(kx - \omega t)$ ,  $\omega$  is, as we have noted many times, the *angular frequency*, of the number of radians per second. The quantity k, which is called by the rather bland name *wave number* is actually the *spacial frequency* or the number of radians per centimeter. When we take the ratio  $\omega/k$ we get

 $\frac{\omega}{k} \frac{\text{radians/second}}{\text{radians/centimeter}} = \frac{\omega}{k} \frac{\text{centimeters}}{\text{second}} (138)$ 

which is clearly a velocity.

As we saw in Chapter 15 of the Physics text and Chapter 2 of the Calculus text,

$$y_1 = A \sin(kx - \omega t)$$
 sine wave moving  
to the right at a speed  $v_{wave} = \omega/k$  (139)

$$y_2 = A \sin(kx + \omega t)$$
 sine wave moving  
to the left at a speed  $v_{wave} = \omega/k$  (140)

If we add  $y_1$  and  $y_2$  we get the standing wave

$$y_1 + y_2 = 2A \operatorname{sinkx} \cos \omega t \quad \frac{standing}{wave}$$
 (141)

You can use the trigonometric identity sin(a+b) = sina cosb + cosa sinb, noting that sin(-b) = -sinb, and cos(-b) = cosb to check Equation (141).

Rather than use the real function  $sin(kx - \omega t)$ , we can, as a trial solution to the wave equation, use the complex function

$$y = Ae^{i(kx - \omega t)}$$
(142)

$$\frac{\partial y}{\partial x} = i k A e^{i (kx - \omega t)}; \quad \frac{\partial y}{\partial t} = (-i\omega) A e^{i (kx - \omega t)}$$
$$\frac{\partial^2 y}{\partial x^2} = (ik)^2 A e^{i (kx - \omega t)} = -k^2 y$$
$$\frac{\partial^2 y}{\partial t^2} = (-i\omega)^2 A e^{i (kx - \omega t)} = -\omega^2 y \quad (143)$$
where  $(-i)^2 = -1$ .

We are now right back to Equation (136) and get the same solution  $v_{wave}^2 = \omega^2/k^2$ . In this case it is actually easier to work with the real function  $\sin(kx - \omega t)$  rather than the complex function  $e^{i(kx - \omega t)}$  because you do not have to take the real part of the complex function at the end. Working with the real variables was not difficult in this case because the wave equation did not mix up sine and cosine functions as the RLC equation did.

For completeness we have

$$y_1 = A e^{i(kx - \omega t)} e^{i(kx$$

$$y_2 = A e^{i(kx + \omega t)} = a e^{i(kx + \omega t)} = a e^{i(kx + \omega t)} e^{i(kx + \omega t)}$$
 (143)

The standing wave solution is

$$y_{\text{standing}} = y_1 + y_2 = A \left[ e^{i (kx - \omega t)} + e^{i (kx + \omega t)} \right]$$
$$= A \left[ e^{i kx} e^{-i \omega t} + e^{i kx} e^{i \omega t} \right]$$
$$= 2A e^{i kx} \left[ \frac{e^{-i \omega t} + e^{i \omega t}}{2} \right]$$
$$= 2A e^{i kx} \cos \omega t$$
$$= 2A (\cos kx + i \sin kx) \cos \omega t$$
$$= 2A \cos kx \cos \omega t + i \left[ 2A \sin kx \cos \omega t \right]$$
(144)

The imaginary part of y<sub>standing</sub> is

$$(y_{\text{standing}})_{\text{imag}} = 2A \sin kx \cos \omega t$$
 (145)

which is the standing wave solution we got using real variables. Using complex variables to get the standing wave solution was not easier than using real variables.

# Calculus 2000-Chapter 6 Introduction to the Schrödinger Wave Equation

In the introduction to Chapter 37 of the Physics text, we quoted the following story from an address by Felix Block to the American Physical Society in 1976.

"Once at the end of a colloquium I heard Debye saying something like: 'Schrödinger, you are not working right now on very important problems...why don't you tell us some time about that thesis of de Broglie, which seems to have attracted some attention?' So in one of the next colloquia, Schrödinger gave a beautifully clear account of how de Broglie associated a wave with a particle, and how he could obtain the quantization rules ... by demanding that an integer number of waves should be fitted along a stationary orbit. When he had finished, Debye casually remarked that he thought this way of talking was rather childish ... To deal properly with waves, one had to have a wave equation." As we mentioned, Schrödinger took Debye's advice, and in the following months devised a wave equation for the electron wave, an equation from which one could calculate the electron energy levels. That wave equation is now the foundation of chemistry.

In this chapter we sketch the ideas that led Schrödinger to formulate an equation involving complex variables to describe the electron. We then go on to solve that equation for the lowest energy spherically symmetric wave functions for the electron in a hydrogen atom. This is enough to show that the Schrödinger equation, without any extra assumptions, is enough to explain the quantized energy levels of hydrogen.

### SCHRÖDINGER'S WAVE EQUATION

Schrödinger's approach to finding a wave equation for the electron was roughly as follows.

De Broglie, suspecting that the electron, like the photon, had a wave nature as well as a particle nature, went back to Einstein's formula for the energy of a photon

$$E = hf$$
(1)

where h is Planck's constant,  $(f = c/\lambda)$  the frequency of the photon and  $\lambda$  its wavelength. Setting  $E = mc^2$ where m is the mass of the photon gives

$$mc^2 = hf = h\frac{c}{\lambda}$$
;  $m = \frac{h}{\lambda c}$ 

Since photons travel at the speed c, the photon's momentum p should be its mass m times its speed c, or

$$p = mc = \left(\frac{h}{\lambda c}\right)c$$

$$p = \frac{h}{\lambda}$$
(2)

Equation (2) is the famous de Broglie formula for the relationship between the wavelength and momentum of any particle. De Broglie explained the quantization of angular momentum in the Bohr theory by assuming that the allowed Bohr orbits were those in which exactly an integral number of wavelengths fit around the orbit.

Schrödinger's job was to find a wave equation based on the two fundamental relationships E = hf for the particle energy and  $p = h/\lambda$  for the particle wavelength. Because we have been writing wave equations in terms of the angular frequency  $\omega$  radians/second rather than the regular frequency f cycles/second, and the wave number (spacial frequency) k radians/cm rather than the wavelength  $\lambda$  cm/cycle, let us first re-express E and p in terms of  $\omega$  and k rather than f and  $\lambda$ . Using dimensions we have

$$f\frac{\text{cycles}}{\text{second}} = \frac{\omega \text{ radians/sec}}{2\pi \text{ radians/cycle}} = \frac{\omega \text{ cycles}}{2\pi \text{ second}}$$
$$\frac{1}{\lambda \text{ cm/cycle}} = \frac{1}{\lambda} \frac{\text{cycles}}{\text{cm}} = \frac{k}{2\pi} \frac{\text{radians/cm}}{\text{radians/cycle}}$$
$$= \frac{k}{2\pi} \frac{\text{cycles}}{\text{cm}}$$
(3)

Using the standard notation

$$\hbar \equiv \frac{h}{2\pi} \qquad h"bar" \tag{4}$$

we get

$$E = hf = h\frac{\omega}{2\pi} = \hbar\omega$$
$$p = \frac{h}{\lambda} = h\frac{k}{2\pi} = \hbar k$$

Thus we get the very simple formulas

$$E = \hbar \omega ; \quad p = \hbar k$$
 (5)

as the relationship between a particle's energy E and momentum p, and its wave's frequency  $\omega$  and wave number k.

Schrödinger's first attempt at finding a wave equation was to start with the relativistic relationship between the energy and momentum of a particle. That relationship, as we saw in the section on particle accelerators, page 28-24 of the Physics text, is

$$E^{2} = p^{2}c^{2} + m_{0}^{2}c^{4} \qquad \begin{array}{c} relativistic \\ relationship \\ between E \ and \ p \end{array}$$
(6)

. . . .

where  $m_0$  is the rest mass of the particle.

To see how to construct a wave equation, let us start with the simple case of a zero rest mass particle, namely the photon. For the photon, we have simply

$$E^{2} = p^{2}c^{2} \qquad \frac{\text{zero rest mass}}{\text{particle}} \tag{7}$$

We will see that the one dimensional wave equation that leads to Equation (7) is

$$\frac{\partial^2 \Psi}{\partial t^2} = c^2 \frac{\partial^2 \Psi}{\partial x^2} \tag{8}$$

where  $\psi$  (psi) is a Greek letter to represent the wave amplitude. (For rubber rope waves  $\psi = y$ , the wave height. For sound waves  $\psi = p$ , the excess pressure.) To check that Equation (8) is the correct equation, use the trial function

$$\Psi = \Psi_0 e^{i (kx - \omega t)} \tag{9}$$

which, as we saw at the end of the last chapter (see Equation 5-142), represents a wave travelling to the right at a speed  $\omega/k$ .

We have

$$\begin{split} \psi &= \psi_0 e^{i (kx - \omega t)} \\ \frac{\partial \psi}{\partial t} &= -i \omega \psi \ ; \quad \frac{\partial^2 \psi}{\partial t^2} = (-i\omega)^2 \psi = -\omega^2 \psi \\ \frac{\partial \psi}{\partial x} &= -i k \psi \ ; \quad \frac{\partial^2 \psi}{\partial x^2} = (-ik)^2 \psi = -k^2 \psi \end{split}$$

Plugging these values into Equation (8) gives

$$\frac{\partial^2 \Psi}{\partial t^2} = c^2 \frac{\partial^2 \Psi}{\partial x^2}$$
(8) repeated  
$$-\omega^2 \Psi = c^2 (-k^2) \Psi$$

The factor  $-\psi$  cancels and we get

$$\omega^2 = c^2 k^2 \tag{10}$$

Multiply through by  $\hbar^2$  and noting that  $E = \hbar \omega$ and  $p = \hbar k$  we get

$$\hbar^2 \omega^2 = c^2 (\hbar^2 k^2)$$
  

$$E^2 = c^2 p^2$$
(11)

which is the result we wanted.

### Exercise 1

For a traveling wave, use the trial function  $\psi = \psi_0 \sin (kx - \omega t)$ 

and show that you get the same result.

You can see that the process is quite straightforward. For each factor of  $\omega$  you want from your differential equation, you put a  $\partial/\partial t$  into the equation. For each factor of (k), you include a  $\partial/\partial x$ .

If we set  $\psi = \vec{E}$  or  $\vec{B}$  in Equation (8) we get the wave equations

$$\frac{\partial^2 \vec{E}}{\partial t^2} = c^2 \frac{\partial^2 \vec{E}}{\partial x^2}$$
(12a)

$$\frac{\partial^2 \vec{\mathbf{B}}}{\partial t^2} = c^2 \frac{\partial^2 \vec{\mathbf{B}}}{\partial x^2}$$
(12b)

These turn out to be the differential form (in one dimension) of the electromagnetic wave we discussed in Chapter 32 in the Physics text. (These are Equations (24a) and (24b) of Chapter 9 of the Calculus text, if we set  $c^2 = 1/\mu_0 \epsilon_0$ .) This should not be surprising, because an electromagnetic wave just represents the wave nature for the zero rest mass photon.

Now that we have some experience constructing wave equations, let us go for the equation for a particle with rest mass. This time let us first convert the relationship between the particle energy E and momentum p into a relationship between  $\omega$  and k. We have

$$E^2 = p^2 c^2 + m_0^2 c^4$$

Setting  $E = \hbar \omega$  and  $p = \hbar k$  gives

$$\hbar^2 \omega^2 = \hbar^2 k^2 c^2 + m_0^2 c^4$$

Dividing through by  $\hbar^2$  gives

$$\omega^2 = c^2 k^2 + \frac{m_0^2 c^4}{\hbar^2}$$
(13)

Using a  $\partial/\partial t$  for each  $\omega$  and a  $\partial/\partial x$  for each k suggests the wave equation

$$\frac{\partial^2 \Psi}{\partial t^2} = c^2 \frac{\partial^2 \Psi}{\partial x^2} - \frac{m_0^2 c^4}{\hbar^2} \Psi$$
(14)

Plugging in the trial solution

$$\psi = \psi_0 e^{i(kx - \omega t)}$$
$$\frac{\partial^2 \psi}{\partial t^2} = -\omega^2 \psi ; \quad \frac{\partial^2 \psi}{\partial x^2} = -k^2 \psi$$

gives

$$-\omega^{2}\psi = -c^{2}k^{2}\psi - \frac{m_{0}^{2}c^{4}}{\hbar^{2}}\psi$$
(15)

cancelling the factor of  $-\psi$  gives

$$\omega^2 = c^2 k^2 + \frac{m_0^2 c^4}{\hbar^2}$$
(16)

which is the result we wanted.

Equation (14) is the one dimensional form of *Schrödinger's relativistic wave equation*. This is the first wave equation Schrödinger found, but he ran into trouble with it.

Consider the case of a particle at rest, or nearly at rest, so that we can neglect  $p^2c^2$  compared to  $m_0^2c^4$ . Then the square of the energy E is approximately equal to the square of the rest energy  $m_0c^2$ 

$$E^2 \approx m_0^2 c^4 \qquad for \\ small p \qquad (17)$$

This equation has two solutions

$$E_{1} = m_{0}c^{2}$$

$$E_{2} = -m_{0}c^{2}$$
(18)

Solution (2) appears to represent a particle with a negative rest energy, a very un-physical thing. The corresponding wave solutions are

$$\Psi_1 = \Psi_0 e^{i(kx - \omega_1 t)}; \quad \hbar \omega_1 = E_1$$
(19)

$$\Psi_2 = \Psi_0 e^{i (kx - \omega_2 t)}$$
;  $\hbar \omega_2 = E_2$  (20)

When you encounter two solutions to a physical problem, and one is nonsense, you usually throw the bad solution out. For example, the hypotenuse of a right triangle is given by the equation

$$c^2 = a^2 + b^2$$
(21)

which has two solutions

$$c_1 = +\sqrt{a^2 + b^2}$$
 (22)

$$c_2 = -\sqrt{a^2 + b^2}$$
(23)

Since you know that you cannot have a negative hypotenuse, you just throw out the un-physical solution  $c_2$ .

Schrödinger tried to throw out the un-physical solution  $\psi_2$  of his relativistic wave equation, but ran into the following problem. If he started with pure  $\psi_1$  waves for the electrons, and let the electrons interact,  $\psi_2$  waves were generated. In other words, if he threw out the un-physical  $\psi_2$  waves, the equations put them back in. We did not have this problem with the Pythagorean theorem.

Schrödinger gave up on the relativistic wave equation and decided to use the nonrelativistic relationship between the kinetic energy E and momentum p of a slowly moving particle. That relationship is

kinetic  
energy E = 
$$\frac{1}{2}mv^2 = \frac{1}{2m}(m^2v^2)$$
 (24)

where v is the speed of the particle, m the rest mass, and mv = p is the momentum. Thus E and p are related nonrelativistically by

$$E = \frac{(mv)^2}{2m} = \frac{p^2}{2m}$$
(25)

Writing  $E = \hbar \omega$ ,  $p = \hbar k$ , the nonrelativistic relationship between  $\omega$  and k is

$$\boxed{\hbar\omega = \frac{\hbar^2 k^2}{2m}} \qquad \begin{array}{c} nonrelativistic \\ relationship \\ between \, \omega \, and \, k \end{array} (26)$$

Schrödinger went to the nonrelativistic form because the relationship  $E = p^2/2m$  does not involve negative rest masses.

To construct a wave equation that gives this nonrelativistic relationship between  $\omega$  and k, we need one time derivative to give the one factor of  $\omega$ , and two x derivatives to give the factor of  $k^2$ . What works, as we will check, is

$$i\hbar \frac{\partial \Psi}{\partial t} = -\frac{\hbar^2}{2m} \frac{\partial^2 \Psi}{\partial x^2} \qquad one dimensional Schrödinger's equation for (27) a free electron$$

With the trial solution

 $\Psi = \Psi_0 e^{i(kx - \omega t)}$ 

$$\frac{\partial \Psi}{\partial t} = -i\omega \Psi$$
;  $\frac{\partial^2 \Psi}{\partial x^2} = -k^2 \Psi$  (28)

we get

$$i\hbar(-i\omega\psi) = \frac{\hbar^2}{2m}k^2\psi$$
$$-i^2\omega\hbar\psi = \frac{\hbar^2}{2m}k^2\psi$$
(29)

The  $\psi$ 's cancel, and with  $-i^2 = 1$ , we are left with the desired result

$$\hbar\omega = \frac{\hbar^2 k^2}{2m}$$
(26) repeated

Equation (27) is the one dimensional form of Schrödinger's equation for a free particle.

In Chapter 2 of the Calculus text, we saw that the equations for rope waves, sound waves, and electromagnetic waves all had second derivatives of both space and time. That is how we got the oscillating solutions. In our study of the RLC circuit, we saw that the presence of a first derivative, the R term in

$$\frac{d^2Q}{dt^2} + \frac{R}{L}\frac{dQ}{dt} + \frac{Q}{LC} = 0$$
 (5-59) repeated

led to an exponential decay.

One might wonder, since there is only a first derivative with respect to time in Schrödinger's equation, shouldn't that lead to an exponential decay with time, of the wave amplitude  $\psi$ ? It did not do so because of the explicit factor of (*i*) in Schrödinger's equation. With the trial solution  $\psi = \psi_0 e^{i (kx - \omega t)}$ the (-*i*) from the first derivative with respect to time was turned into a 1 by the *i* in the  $\partial/\partial t$  term. Thus by having an (*i*) in Schrödinger's equation itself, we can get an oscillating solution with a first time derivative.

The reason we have introduced Schrödinger's equation after a chapter on complex variables is that factor of (i) in the equation itself. With the other differential equations we have discussed so far, we had the choice of using real or complex variables. But we cannot write, let alone solve, Schrödinger's equation without the use of complex variables.

### Exercise 2

In three dimensions, the momentum vector  $\vec{p} = (p_X, p_y, p_z)$  has a magnitude p given by the Pythagorean theorem as

$$p^{2} = (p_{X}^{2} + p_{y}^{2} + p_{z}^{2})$$
(30)

With  $\vec{p} = \hbar \vec{k}$ , we have

$$p^{2} = \hbar^{2} (k_{x}^{2} + k_{y}^{2} + k_{z}^{2})$$
(31)

We got the one dimensional wave equation by replacing  $k_x^2$  by  $\partial^2/\partial x^2$ . This suggests that the extension of Equation (27) to describe three dimensional plane waves should be

$$i\hbar\frac{\partial\psi}{\partial t} = -\frac{\hbar^2}{2m} \left( \frac{\partial^2\psi}{\partial x^2} + \frac{\partial^2\psi}{\partial y^2} + \frac{\partial^2\psi}{\partial z^2} \right)$$
(32)

As a trial solution, try the guess

$$\psi = e^{i(\vec{k}\cdot\vec{x}-\omega t)} = e^{i(k_xx+k_yy+k_zz-\omega t)}$$
(33)

and show that the guess implies

$$\hbar \omega = \frac{\hbar^2}{2m} (k_x^2 + k_y^2 + k_z^2)$$
(34)

and

$$\mathsf{E} = \frac{\mathsf{p}^2}{2\mathsf{m}}$$

## POTENTIAL ENERGY & SCHRÖDINGER'S EQUATION

The relationship  $E = p^2/2m = mv^2/2$  is for a free particle traveling at a constant speed v. If the particle has a potential energy V( $\vec{x}$ ), like spring potential energy

$$V(x) = -\frac{1}{2}Kx^{2}$$
spring
potential
energy
(35)

where K is the spring constant, then the formula for the total nonrelativistic energy E is

$$E = \frac{1}{2}mv^{2} + V(x) = \frac{p^{2}}{2m} + V(x)$$
(36)

In terms of  $\omega$  and k we have

$$\hbar\omega = \frac{\hbar^2 k^2}{2m} + V(x)$$
(37)

and the corresponding one dimensional wave equation should be

$$i\hbar\frac{\partial\psi}{\partial t} = -\frac{\hbar^2}{2m}\frac{\partial^2\psi}{\partial x^2} + V(x)\psi$$
one
dimensional
Schrödinger
equation
(38)

If you did Exercise (2), it is clear that the three dimensional form of Schrödinger's equation is expected to be

$$i\hbar\frac{\partial\Psi}{\partial t} = -\frac{\hbar^2}{2m}\left(\frac{\partial^2\Psi}{\partial x^2} + \frac{\partial^2\Psi}{\partial y^2} + \frac{\partial^2\Psi}{\partial z^2}\right) + V(x,y,z)\Psi$$
(39)

In Chapter 4 of the Calculus text, we discussed the combination of derivatives  $\partial^2/\partial x^2 + \partial^2/\partial y^2 + \partial^2/\partial z^2$  and gave them the special name

$$\nabla^2 \equiv \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2} \qquad definition \qquad (40)$$

With this notation, the three dimensional form of Schrödinger's equation can be written in the more compact and familiar form



We can immediately get back to the one dimensional Schrödinger's equation by replacing  $\nabla^2$  by  $\partial^2/\partial x^2$ .

### THE HYDROGEN ATOM

The reason Schrödinger developed his wave equation was to handle the electron waves in hydrogen in a mathematically rigorous way. To apply Schrödinger's equation of the hydrogen atom, you use the fact that the electron is bound to the proton nucleus by a Coulomb force of magnitude  $e^2/r^2$  whose potential V(r) is

$$V(\mathbf{r}) = \frac{-e^2}{\mathbf{r}} \qquad \begin{array}{c} Coulomb\\ potential\\ energy \end{array}$$
(42)

With this potential energy, Schrödinger's equation (41) for the hydrogen atom becomes

$$\frac{\hbar}{i}\frac{\partial\psi}{\partial t} = \frac{\hbar^2}{2m}\nabla^2\psi - \frac{e^2}{r}\psi$$
Schrödinger's
equation for
hydrogen atom
(43)

Solving Equation (43) is not easy. The first problem we encounter is the fact that we have been writing  $\nabla^2 = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}$  using Cartesian coordinates x, y, z, while the Coulomb potential –  $\frac{e^2}{r}$ has spherical symmetry. The best way to handle the situation is to use a coordinate system that has the same symmetry as the potential energy.

The coordinate system of choice is the *spherical polar coordinate system* that has an inherent spherical symmetry. This coordinate system is described in Chapter 4 of the Calculus text and indicated in Figure (1). Instead of locating a point by giving its x, y, and z coordinates, we locate it by the r,  $\theta$  and  $\phi$  coordinates. The quantity r is the distance from the origin,  $\theta$  the angle down from the z axis, and  $\phi$  the angle over from the x axis, as shown.



**Figure 1** Spherical polar coordinates.

In the appendix to Chapter 4 of the Calculus text, we calculated  $\nabla^2$  in spherical polar coordinates. The result was

$$\nabla^{2} \Psi = \frac{1}{r} \frac{\partial^{2}}{\partial r^{2}} (r \Psi) + \frac{1}{r^{2} \sin \theta} \frac{\partial}{\partial \theta} \left( \sin \theta \frac{\partial \Psi}{\partial \theta} \right) + \frac{1}{r^{2} \sin \theta} \frac{\partial^{2} \Psi}{\partial \phi^{2}}$$
(44)

This surely does not look simpler than  $\nabla^2 = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}$ , but it does allow you to find solutions to Schrödinger's equation for the hydrogen atom.

In the appendix to this chapter, we calculate some spherically symmetric solutions to Schrödinger's equations. These are solutions that depend only on r, namely  $\psi = \psi(r)$ , so that  $\partial \psi / \partial \theta = 0$  and  $\partial \psi / \partial \phi = 0$ , which eliminates the second and third terms in Equation (44). The solutions we get, (we solve one and leave the second as a homework exercise) are

$$\Psi_1 = e^{-r/a_0} e^{-i\omega_1 t} \tag{45}$$

$$\Psi_{2} = \left(1 - \frac{r}{2a_{0}}\right) e^{-r/2a_{0}} e^{-i\omega_{2}t}$$
(46)

where  $a_0$  has the value

$$a_0 = \frac{\hbar^2}{me^2} \qquad Bohr \, radius \tag{47}$$

This quantity  $a_0$  is the **Bohr radius**, the radius of the smallest orbit in the Bohr theory of hydrogen. (See Exercise 7 in Chapter 35 of the Physics text.)

### **Exercise 3**

Go to Appendix II of this chapter (page 6-14) and study the steps that led to the solution  $\psi_1$ . Then work Exercise 5 to find the solution  $\psi_2$ . After that return here and continue reading.

A special feature we discover when we solve Schrödinger's equation in Appendix II, is that in order for  $\psi_1$  and  $\psi_2$  to be solutions of Schrödinger's equation (43), *the frequencies*  $\omega_1$  *and*  $\omega_2$  *have to have the following values* 

$$\hbar\omega_1 = \frac{-e^4m}{2\hbar^2} = -13.6\,\text{eV}$$
(48)

$$\hbar\omega_2 = \frac{-e^4m}{8\hbar^2} = -3.60\,\text{eV}$$
(49)

You can immediately see that  $\hbar\omega_1$  is the energy of the electron in the lowest hydrogen energy level, and  $\hbar\omega_2$  is the electron energy in the second energy level. Just looking at the spherically symmetric solutions begins to tell us that Schrödinger's equation is going to explain, in a natural way, the hydrogen energy levels.

As we mentioned in our discussion of the hydrogen atom in Chapter 38 of the Physics text, there are many allowed standing wave patterns for the electron in hydrogen. In Figure (38-1), reproduced on the next page, we show sketches of the six lowest energy patterns  $\psi_{n,\ell,m}$  labeled by their energy quantum number (n), angular momentum quantum number ( $\ell$ ) and z projection of angular momentum quantum number (m). We noted that all the zero angular momentum patterns ( $\ell = 0$ ) are spherically symmetric. By solving Schrödinger's equation for spherically symmetric standing waves, we began to generate the  $\ell = 0$  patterns. Explicitly, the waves we got are

$$\psi_{1,0,0} = \psi_1 \text{ (of Equation 45)}$$
$$\psi_{2,0,0} = \psi_2 \text{ (of Equation 46)}$$

To solve for the non symmetric patterns like  $\psi_{2,1,1}$  that have angular momentum, you have to be able to handle angular terms involving  $\theta$  and  $\phi$  in the formula (44) for  $\nabla^2$ . Differential equations involving  $\nabla^2$  have been studied for well over a century, and the angular terms, which are common to many of these equations, have been carefully worked out with standardized notation. The angular dependence of the non spherical standing waves involve what are called *spherical harmonics* which are briefly discussed in Appendix II of this chapter.



There are 8 more $n = 3$ patterns
in addition to the one shown.
The $\ell$ and m quantum numbers
are $\emptyset = 1$ · m = 1 0 1
$\ell = 1,  m = 1,  0, -1$ $\ell = 2;  m = 2,  1,  0, -1, -2.$

 $n = 3, \ \ell = 0, \ m = 0$ 



 $n = 1, \ \ell = 0, \ m = 0$ 

### Figure 38-1 (page 38-3 of the Physics text)

(a)

The lowest energy standing wave patterns in hydrogen. The intensity is what you would see looking through the wave. We have labeled  $\Psi_1$  and  $\Psi_2$  on the diagram.

### INTERPRETATION OF SOLUTIONS TO SCHRÖDINGER'S EQUATION

Bohr's theory of the hydrogen atom, although quite successful, was based on Newtonian mechanics with the ad hoc assumption that angular momentum was quantized in units of  $\hbar$ . De Broglie's theory suggested that the reason for the quantization of angular momentum was due to the wave nature of the electron, but he also treated the electron wave in a rather ad hoc manner. If one assumes that Schrödinger's equation rather than Newtonian mechanics provides the basic theory for the electron in hydrogen, then all the quantized energy levels follow a direct consequence of the theory. No extra assumptions have to be fed in. Schrödinger had found the theory to replace Newtonian mechanics in describing atoms.

But questions remained. The electron's wave nature was well established, but what was the meaning of the electron wave? The answer to that was provided a couple of years later by Max Born, who was calculating how electron waves would be scattered by atoms. The calculations suggested to him that the electron wave should be interpreted as a probability wave, as we discussed in Chapter 40 of the Physics text.

One of the main features of a probability wave is that it has to be represented by a real, positive number. You cannot have negative probabilities or imaginary probabilities. But so far, our electron waves are described by a complex variable  $\psi$ , obtained from an equation that was itself complex. How do we get real positive numbers from the complex  $\psi$ ?

We ran into a somewhat similar problem in our discussion of electromagnetic radiation. Maxwell's equations predict that light waves consist of electric and magnetic fields  $\vec{E}$  and  $\vec{B}$ . Yet most of the time we are concerned with the intensity or energy density of a light wave. To predict the intensity from Maxwell's theory, we have to know how to calculate the intensity from the vectors  $\vec{E}$  and  $\vec{B}$ . The answer is that the intensity is proportional to the square of  $\vec{E}$  and  $\vec{B}$ . If we use the correct units, the intensity is proportional to  $(\vec{E} \cdot \vec{E} + \vec{B} \cdot \vec{B})$ . These dot products  $\vec{E} \cdot \vec{E}$  and  $\vec{B} \cdot \vec{B}$  are always positive numbers and therefore can represent an energy density or intensity.

If we can get a positive number for a vector field by taking the dot product of the vector with itself, what do we do to get a positive number from a complex  $\psi$ ? The answer, as we mentioned at the beginning of Chapter 5 (see Equation 5-26), is that we get a real positive number from a complex number by multiplying by the *complex conjugate*. To remind you how this works, suppose that we have separated  $\psi$  into its real and imaginary parts

$$\Psi = \Psi_{\text{real}} + i\Psi_{\text{imag}} \tag{50}$$

where both  $\psi_{real}$  and  $\psi_{imag}$  are real numbers. Then the complex conjugate, which we designate by  $\psi^*$ , is defined by changing (*i*) to (-*i*)

$$\Psi^* = \Psi_{\text{real}} - i\Psi_{\text{imag}} \tag{51}$$

To calculate the complex conjugate  $\Psi^*$  you do not have to separate the function into real and imaginary parts ahead of time. You get the same result by replacing all (*i*) by (-*i*) in the complex formula.

When you multiply a complex number  $\psi$  by its complex conjugate  $\psi^*$ , the result is a real positive number, as you can see below

$$\psi^* \psi = (\psi_{real} - i\psi_{imag}) (\psi_{real} + i\psi_{imag})$$
$$= \psi_{real} \psi_{real} + i\psi_{real} \psi_{imag}$$
$$- i\psi_{imag} \psi_{real} - i^2 \psi_{imag} \psi_{imag}$$

The  $i\psi_{\text{real}}\psi_{\text{imag}}$  terms cancel, and with  $-i^2 = 1$  we get

$$\Psi^* \Psi = \Psi_{\text{real}}^2 + \Psi_{\text{imag}}^2 \tag{52}$$

and thus  $\psi^* \psi$  is a real, positive number.

For electron waves, the positive number  $\Psi^* \Psi$  represents the intensity of the wave in much the same way that  $(\vec{E} \cdot \vec{E} + \vec{B} \cdot \vec{B})$  represented the intensity of the electromagnetic wave.

### Normalization

In describing probabilities, one usually represents a probability of 1 as being certainty, and that the probability of an event as being allowed to range from zero to one. If the wave function  $\psi$  is to represent a probability wave for an electron, we have to include the idea that the probability of something ranges from zero to one.

The intensity  $\psi^*\psi$  is a density that varies over space. If you have an energy density, call it *E*, then the total energy *E* is the integral over all of space of the energy density *E*. We can write this symbolically as

$$E = \int_{\text{all space}} E(x,y,z) d^{3}V$$
 (53)

where, if we are using Cartesian coordinates, the volume element  $d^3V$  would be  $(dx \times dy \times dz)$ .

If we are to interpret  $\psi^* \psi$  as a probability density, then the total probability should be the integral of the probability density over all space. We can write this as

total  
probability = 
$$\int_{\text{all space}} \psi^* \psi \, d^3 V$$
 (54)

The question is, this is the total probability of what? If we are talking about the electron wave in hydrogen, and we think of  $\psi^* \psi d^3 V$  as the probability of finding the electron in some small volume element  $d^3 V$ , then if we sum these probabilities over all space, we should end up with the total probability of finding the electron somewhere in space. If the hydrogen atom has one electron, and you look everywhere, you should eventually find the electron with a probability (1). Thus the total probability should be given by the formula

$$1 = \int_{\text{all space}} \psi^* \psi \, d^3 V$$
 (55)

The wave functions  $\psi_1$  and  $\psi_2$  that we presented you in Equations (45) and (46) do not have this property.

Let us see what the integral of  $\psi_1^*\psi_1$  over all space is. We have

$$\Psi_1 = e^{-r/a_0} e^{-i\omega_1 t}$$
 (56a)

$$\Psi_1^* = e^{-r/a_0} e^{+i\omega_1 t}$$
 (change - *i* to *i*) (56b)

so that

$$\Psi_{1}^{*}\Psi_{1} = e^{-r/a_{0}}e^{+i\omega_{1}t} e^{-r/a_{0}}e^{-i\omega_{1}t}$$

$$\psi_{1}^{*}\Psi_{1} = e^{-2r/a_{0}}$$
(57)

The  $e^{i\omega_1 t's}$  cancelled and we end up with a real positive density.

To integrate  $\psi^*\psi$  over all space, we notice that since  $\psi^*\psi$  is spherically symmetric, we can take  $d^3V$  as the volume of the spherical shell shown in Figure (2), a shell of radius r and thickness dr. That volume is

$$d^{3}V = (4\pi r^{2})dr$$
 (58)

because  $4\pi r^2$  is the area of a sphere of radius r. Throughout the shell,  $\psi^*\psi$  has the same value  $e^{-2r/a_0}$ , thus our volume integral is simply

$$\int_{\text{all space}} \psi^* \psi \, d^3 V = \int_{r=0}^{\infty} e^{-2r/a_0} (4\pi r^2) dr$$
 (59)

Being somewhat lazy, we look up in our short table of integrals, the integral of  $r^2e^{-\alpha r}$ . After some manipulation shown in Appendix 1, we get

$$4\pi \int_{0}^{\infty} r^2 e^{-2r/a} 0 dr = \pi (a_0)^3$$
 (60)

The result is that the integral of  $\psi^* \psi$  over all space is  $\pi(a_0)^3$  instead of the desired value of 1.





We can use as the volume element  $d^{3}V$  the spherical shell of radius r and thickness dr.

To fix this problem, we use a so-called normalized wave function  $(\psi_1)_{normalized}$ , which is simply  $\psi_1$  multiplied by an appropriate normalization constant C. To find out what C should be, write

$$(\Psi_1)_{\text{normalized}} = C\Psi_1 \tag{61a}$$

$$(\psi_1^*)_{\text{normalized}} = C^* \psi_1^* \tag{61b}$$

where, if we want, the normalization constant can be complex. Then we have

$$(\Psi_{1}^{*})_{\text{normalized}}(\Psi_{1})_{\text{normalized}} = (C^{*}C) \Psi_{1}^{*}\Psi_{1}$$

$$1 = \int_{\text{all space}} (\Psi_{1}^{*})_{\text{normalized}}(\Psi_{1})_{\text{normalized}} d^{3}V$$

$$= C^{*}C \int_{\text{all space}} \Psi^{*}\Psi d^{3}V \qquad (62)$$

$$= C^*C\pi(a_0)^3$$

Thus

$$C^*C = \frac{1}{\pi(a_0)^3}$$
(63)

The simplest choice is to take C real, giving

$$C = \frac{1}{\sqrt{\pi(a_0)^3}} \begin{bmatrix} normalization \\ constant for \ \psi_l \end{bmatrix}$$
(64)

As a result our normalized wave function becomes

$$(\Psi_1)_{\text{normalized}} = \frac{1}{\sqrt{\pi(a_0)^3}} e^{-2r/a_0} e^{-i\omega t}$$
  
(65)

When you look at tables of wave functions, you will see factors like  $1/\sqrt{\pi(a_0)^3}$  or  $\sqrt{3/8\pi}$ . They are merely the normalization constants. In one sense, the normalization constants just make the formulas look complicated. Most of the physics in our equation for  $\psi_1$  is contained in the factor  $e^{-r/a_0}$ . It tells us that the electron wave decays exponentially as we go out from the proton, decaying by a factor of 1/e when we go out one Bohr radius  $a_0$ . The intensity, or probability  $\psi^* \psi$  is proportional to  $e^{-2r/a_0}$  and thus drops off by a factor  $1/e^2$  when we are a Bohr radius from the proton. We also calculated the energy levels  $E_1$  and  $E_2$  without worrying about the normalization constants. It is nice to have a table that gives you the normalization constants, but you get a better insight into the shape of the standing wave patterns if you have another table without them.

### **Exercise 4**

At what finite radius is there zero probability of finding an electron when the electron is in the n = 2,  $\ell = 0$ , m = 0 standing wave pattern? Explain why and sketch the intensity  $\psi_{2,0,0}^*\psi_{2,0,0}$ 

### THE DIRAC EQUATION

Our story is incomplete if we stop our discussion of particle wave equations with Schrödinger's equation. As successful as that equation is, it still does not handle relativistic effects. As we saw, Schrödinger could avoid the negative rest mass solutions by starting with the nonrelativistic formula  $E = p^2/2m$  rather than the relativistic one  $E^2 = p^2c^2 + m_0^2c^4$ .

It appeared to Dirac that the reason Schrödinger could avoid the nonphysical solutions is because the nonrelativistic equation involves only the first derivative with respect to time  $\partial \psi / \partial t$ , rather than the second derivative  $\partial^2 \psi / \partial t^2$  that appeared in the relativistic equation (see Equation (14). Dirac thought that if he could develop a relativistic wave equation that avoided second time derivatives, then perhaps he could avoid the un-physical negative mass solutions.

By 1929, when Dirac was working on the problem, it was known that the electron had two spin states, spin up and spin down. It was these two spin states, along with the Pauli exclusion principle, that led to an understanding of the structure of the periodic table. These spin states are not included in or explained by Schrödinger's equation.

Slightly earlier, Wolfgang Pauli had introduced a new mathematical quantity called a *spinor* to describe the spin state of the electron. Spinors are quantities, involving complex numbers, that are in a sense half way between a scalar number and a vector. The existence of such a mathematical quantity was unknown until its invention was required to explain the electron. Pauli was able to modify Schrödinger's equation with the use of spinors to include the effects of electron spin. Dirac found that by using a certain combination of spinors, he could write a relativistic wave equation for the electron that had only a first order time derivative  $\partial \psi / \partial t$ . He hoped that this equation would avoid the un-physical negative mass solutions.

Dirac's equation was successful in that it not only included all the results of Schrödinger's and Pauli's equations, but it also correctly predicted tiny relativistic effects that could be detected in the spectra of hydrogen. However, Dirac soon found that his equation also led to the apparently negative mass solutions.

Dirac could not throw his equation away because it successfully predicted relativistic effects that were observed by experiment. Instead he found a new interpretation of the previously undesirable solutions. He found that these solutions could be reinterpreted as the wave for a particle whose mass was positive but whose electric charge was of the opposite sign. The equation led to the prediction that there should exist a particle with the same rest mass as the electron but with a positive electric charge. That particle was observed four years later in Carl Anderson's cloud chamber in the basement of the physics building at Caltech. It became known as the **positron**.

We now know that any relativistic wave equation for a particle has two kinds of waves for a solution. One represents matter particles, and the other, like the wave for the positron represents *antimatter*. If you have a relativistic wave equation, even if you start only with matter particles, the equation contains the mechanism for particle-antiparticle pair creation. You let the matter particles interact, and antimatter has a finite probability of being created. That is why Schrödinger and Dirac could not suppress the antimatter waves in the relativistic equations. However, by going to a nonrelativistic equation, representing situations where not enough energy is available to create electron positron pairs, Schrödinger could avoid the antimatter waves.

## **Appendix I – Evaluation of a Normalization Integral**

Our normalization integral is

$$\int_{\text{all space}} \psi^* \psi \, d^3 V = 4\pi \int_{r=0}^{\infty} r^2 e^{-2r/a_0} \, dr \quad (59) \text{ repeat}$$

Looking for the integral of  $r^2e^{-\alpha r}$  in our short table of integrals in the formulary, we find instead

$$\int x^2 e^{-ax} dx = \frac{1}{a^3} (a^2 x^2 + 2ax + 2) e^{-ax} \quad (66)$$

If we set x = r and integrate from 0 to infinity, we have

$$\int_{0}^{\infty} r^{2} e^{-ar} dr = \frac{1}{a^{3}} (a^{2}r^{2} + 2ar + 2)e^{-ar} \Big|_{0}^{\infty}$$
$$= \frac{1}{a^{3}} (a^{2}R^{2} + 2aR + 2)e^{-aR} \Big|_{R=\infty}$$
$$- \frac{1}{a^{3}} (a^{2}0^{2} + 2a \times 0 + 2)e^{-a \times 0}$$
(67)

The exponential decay is so powerful that in the limit of large R, a term of the form  $R^n e^{-aR}$  goes to zero for any value of n for positive (a). Thus all terms with a  $e^{-aR}$  go to 0 as R goes to infinity. With  $e^0 = 1$ , we are left with

$$\int_{0}^{\infty} r^2 e^{-ar} dr = \frac{2}{a^3}$$
(68)

Now set  $a = 1/2a_0$  and we get

$$4\pi \int_{0}^{\infty} r^{2} e^{-2r/a_{0}} dr = 4\pi \frac{2}{(2/a_{0})^{3}}$$
$$= \pi (a_{0})^{3}$$
(69)

### **APPENDIX II - An introduction to Schrödinger's Equation Applied to the Hydrogen Atom**

### **The Hydrogen Atom**

Schrödinger's first major success with his wave equation was to solve for the electron standing waves in hydrogen, and to determine the electron energies in each of the standing wave patterns. For an electron in hydrogen, the potential energy is given by Coulomb's law as

$$V(r) = -\frac{e^2}{r}$$
 (42) repeated

where –e is the charge on the electron and r is the separation of the electron and proton. Thus the equation Schrödinger had to solve for hydrogen is the three dimensional equation

$$i\hbar\frac{\partial\psi}{\partial t} = -\frac{\hbar^2}{2m}\nabla^2\psi - \frac{e^2}{r}\psi$$
Schrödinger's  
equation for  
hydrogen atom

(43) repeated

Quite a few steps are required to obtain solutions to Equation (43). The first is to look for solutions of definite frequency  $\omega$  or energy  $E = \hbar \omega$  by using the trial function

$$\begin{aligned} \Psi &= \Psi(\mathbf{x}, \mathbf{y}, \mathbf{z}) \, \mathrm{e}^{-\,\mathrm{i}\omega t} \\ &= \Psi(\mathbf{x}) \, \mathrm{e}^{-\,\mathrm{i}\omega t} \end{aligned} \tag{70}$$

where we will use the bold face  $\mathbf{x}$  to stand for (x,y,z). Plugging this guess into Equation (43) gives

$$i\hbar(-i\omega)\psi(\mathbf{x})e^{-i\omega t} = -\frac{\hbar^2}{2m} \left[\nabla^2 \psi(\mathbf{x})\right]e^{-i\omega t}$$
$$-\frac{e^2}{r}\psi(\mathbf{x})e^{-i\omega t}$$

The factor  $e^{-i\omega t}$  cancels and we are left with

$$\hbar\omega\psi(\mathbf{x}) = -\frac{\hbar^2}{2m} \left[\nabla^2\psi(\mathbf{x})\right] - \frac{e^2}{r}\psi(\mathbf{x})$$
(71)

With  $\hbar \omega = E$ , this becomes

$$\mathbf{E}\psi(\mathbf{x}) = -\frac{\hbar^2}{2m} \left[\nabla^2 \psi(\mathbf{x})\right] - \frac{e^2}{r} \psi(\mathbf{x})$$
(72)

The next step is to note that it is not convenient to handle a spherically symmetric potential  $V(r) = -e^2/r$  using Cartesian coordinates x, y, and z. In the Chapter 4 of the Calculus text we derived the formula for  $\nabla^2$  in *spherical polar coordinates* r,  $\theta$ ,  $\phi$  which are shown in Figure (1) reproduced here. In these spherical coordinates we show, after considerable work, that  $\nabla^2 \psi$  is given by Equation (4-10) as

$$\nabla^{2} \Psi = \frac{1}{r} \frac{\partial^{2}}{\partial r^{2}} (r \Psi)$$
  
+ 
$$\frac{1}{r^{2} \sin \theta} \frac{\partial}{\partial \theta} \left( \sin \theta \frac{\partial \Psi}{\partial \theta} \right)$$
  
+ 
$$\frac{1}{r^{2} \sin^{2} \theta} \frac{\partial^{2} \Psi}{\partial \phi^{2}}$$
(4-10)

(Note: many texts write the first term as  $1/r^2 \partial/\partial r (r^2 \partial \psi/\partial r)$  which is an equivalent but usually less convenient form.)



**Figure 1 (repeated)** Spherical polar coordinates.

If we look at only the spherically symmetric solutions where

$$\psi(\mathbf{x},\mathbf{y},\mathbf{z}) = \psi(\mathbf{r}) \qquad \begin{array}{c} spherically\\ symmetric\\ wave \end{array} \tag{73}$$

then  $\partial \psi(r) / \partial \theta = 0$ ,  $\partial \psi / \partial \phi = 0$ , and only the radial part of  $\nabla^2 \psi(r)$  survives. Schrödinger's equation for the spherically symmetric waves of energy E becomes

$$E\psi = -\frac{\hbar^2}{2m}\frac{1}{r}\frac{\partial^2}{\partial r^2}(r\psi) - \frac{e^2}{r}\psi$$
(74)

Multiplying through by  $2mr/\hbar^2$ , Equation (74) can be written in the form

$$\frac{\partial^2}{\partial r^2}(r\psi) + \left(\frac{a}{r} + b\right)r\psi = 0$$
(75)

where

$$a = \frac{2me^2}{\hbar^2} ; \quad b = \frac{2mE}{\hbar^2}$$
(76)

If we define the variable u(r) by

$$\mathbf{u} = \mathbf{r}\boldsymbol{\Psi}; \quad \boldsymbol{\Psi} = \frac{\mathbf{u}}{\mathbf{r}} \tag{77}$$

our equation for u becomes

$$\frac{\partial^2 \mathbf{u}}{\partial r^2} + \left(\frac{\mathbf{a}}{r} + \mathbf{b}\right)\mathbf{u} = 0 \tag{78}$$

### Exercise 5

Derive Equation (78) starting from Equation (74).

Equation (78) is a differential equation we have not encountered before. Neither of our familiar guesses for a solution, like  $u = e^{-\alpha r}$  or  $u = \sin\omega r$ , will work, as you can check for yourself. What does work is the function we will call  $u_1$ , which is

$$u_1(r) = r e^{-\alpha r} \qquad guess \qquad (79)$$

Plugging our guess into Equation (79) gives

$$\frac{du_1}{dr} = e^{-\alpha r} - \alpha r e^{-\alpha r}$$
$$\frac{d^2 u_1}{dr^2} = -\alpha e^{-\alpha r} - \alpha e^{-\alpha r} + \alpha^2 r e^{-\alpha r}$$
$$Thus \frac{d^2 u_1}{dr^2} + \left(\frac{a}{r} + b\right)u = 0 \text{ becomes}$$
$$-2\alpha e^{-\alpha r} + \alpha^2 r e^{-\alpha r} + \frac{a}{r} r e^{-\alpha r} + br e^{-\alpha r} = 0$$

The common factor  $e^{-\alpha r}$  cancels and we are left with

$$\left[-2\alpha + a\right] + r\left[\alpha^2 + b\right] = 0 \tag{80}$$

The only way we can satisfy Equation (80) for arbitrary values of r is to set both square brackets separately equal to zero, giving

$$2\alpha = a ; \quad \alpha = a/2 \tag{81a}$$

$$\alpha^2 = -b \tag{81b}$$

Squaring Equation (81a) gives

$$\alpha^2 = \frac{a^2}{4} \tag{81c}$$

For Equations (81b) and (81c) to be consistent, the constants (a) and (b) must satisfy the relationship

$$-b = \frac{a^2}{4} \tag{82}$$

To see what Equation (82) implies, let us put back in the values of (a) and (b)

$$a = \frac{2me^2}{\hbar^2}$$
;  $\frac{a^2}{4} = \frac{1}{4} \times \frac{4m^2e^4}{\hbar^4}$  (83a)

$$-b = \frac{-2mE}{\hbar^2}$$
(83b)

Thus Equation (82) requires

$$-\frac{2mE}{\hbar^2} = \frac{m^2e^4}{\hbar^4}$$

or

$$E = -\frac{me^4}{2\hbar^2} = -13.6 \text{ eV}$$
 (84)

In our study of the Bohr theory, we found that the lowest energy level of the hydrogen atom was  $E_1 = -me^4/2\hbar^2$  which turns out to be -13.6 electron volts. We now see that if the hydrogen wave amplitude is given by the solution  $u_1$ , or  $\psi_1 = u_1 r$ , then the energy of the electron in this wave pattern must be the same as the lowest energy level of the Bohr theory. This is a prediction of Schrödinger's wave equation without any arbitrary added assumptions like assuming angular momentum is quantized.

To see what the wave pattern is that corresponds to the energy level  $E_1$ , note that the Bohr radius  $a_0$ , the radius of the smallest Bohr orbit in the Bohr theory, is given by

$$a_0 = \frac{\hbar^2}{me^2} \qquad Bohr \ radius \tag{85}$$

Thus our constant (a) in Equation (77) can be written

$$a = \frac{2me^2}{\hbar^2} = \frac{2}{a_0}$$
(86)

Thus Equation (81a) requires that

$$a = 2\alpha = \frac{2}{a_0}; \quad \alpha = \frac{1}{a_0}$$
 (87)

and the wave function  $\psi_1(\mathbf{r})$  is given by

$$\psi_{1}(r) = \frac{u_{1}(r)}{r} = \frac{r e^{-\alpha r}}{r} = e^{-\alpha r}$$

$$\psi_{1}(r) = e^{-r/a_{0}}$$
(88)

The electron wave decays exponentially as we go out from the nucleus, decaying by a factor of 1/e when we go out one Bohr radius. We have just used Schrödinger's equation to solve for the ground state wave function, the lowest energy level standing wave pattern in hydrogen.

### **The Second Energy Level**

In the following exercise you will find another spherically symmetric solution for the hydrogen atom.

### **Exercise 6**

Try the guess

$$u_2(r) = (r + cr^2)e^{-\alpha r}$$
,  $u_2 = r\psi_2$  (89)

as a possible solution to Equation (78) where (c) is an unknown constant. Show that for (89) to be a solution, you have to satisfy the conditions

$$-2\alpha + 2c + a = 0$$
 (90a)

$$\alpha^2 - 4c\alpha + ac + b = 0 \tag{90b}$$

$$\alpha^2 c + bc = 0 \tag{90c}$$

Then show that this requires  $\alpha^2 = -b$  as before, and that

$$-b = \frac{a^2}{16} \implies \frac{-2mE_2}{\hbar^2} = \frac{1}{16} \times \frac{4m^2e^4}{\hbar^4}$$
 (91)

or

$$\mathsf{E}_2 = -\frac{1}{4} \left( \frac{\mathsf{me}^4}{2\hbar^2} \right) = \frac{-13.6 \,\mathrm{eV}}{4} = -3.60 \,\mathrm{eV} \qquad (92)$$

Then show that  $\psi_2(r)$  is given by

$$\Psi_2 = (1 - \frac{r}{2a_0})e^{-r/2a_0}$$

$$E_2 = -3.6 \text{ eV}$$
(93)

In the Bohr theory, the energy levels  $E_n$  are given by

$$E_n = \frac{E_1}{n^2} = -\frac{13.6 \text{ eV}}{n^2}$$
(94)

The second energy level  $E_2$  is thus

$$E_2 = \frac{E_1}{(2)^2} = \frac{E_1}{4} = -3.6 \text{ eV}$$

Thus the wave pattern you solved for in Exercise (8) is the spherically symmetric standing wave pattern in the second energy level. It is what we have called the n = 2,  $\ell = 0$  wave pattern. Note that in the solution

$$\psi_2(\mathbf{r}) = \left(1 - \frac{\mathbf{r}}{2a_0}\right) e^{-\mathbf{r}/2a_0}$$
 (93) repeated

when we are at a distance

$$\frac{\mathbf{r}}{2\mathbf{a}_0} = 1$$
;  $\mathbf{r} = 2\mathbf{a}_0$  (95)

the wave pattern in Equation (93) goes to zero. This means that the standing wave  $\psi_2(r)$  has a spherical node out at a distance  $r = 2a_0$ . This is the spherical node we saw in the  $\psi(n = 2, \ell = 0)$  pattern shown in the Physics text, Figure (38-1) repeated here.



**Figure 38-1a** Hydrogen atom standing wave pattern for  $n = 2, \ell = 0.$ 

Figure 3

Tacoma Narrows bridge in an n = 2second harmonic standing wave pattern.



(Movie. Press esc to stop)

If you try a guess of the form

$$u_3(r) = (1 + c_2 r + c_3 r^2) e^{-\alpha r}$$
(96)

you end up with a spherical wave pattern  $\psi_3(r)$  that has two spherical nodes, and has an energy

$$E_3 = \frac{E_1}{3^2}$$
 (97)

which is the third energy level.

You can now see the pattern. We can generate all the spherically symmetric  $\ell = 0$  wave patterns by adding terms like  $c_4 r^3$ ,  $c_5 r^4$ ,  $\cdots c_n r^{n-1}$  to our guess for  $u_n(r)$ . Solving for all the constants, we end up with

$$\mathbf{E}_{\mathbf{n}} = \frac{\mathbf{E}_1}{\mathbf{n}^2} \tag{98}$$

which is the energy level structure Bohr discovered.



**Figure 38-1i** Wave pattern for n = 3, l = 0.
## Non Spherically Symmetric Solutions

It was fairly easy to handle the spherically symmetric solutions to Schrödinger's equation for hydrogen, because we did not have to deal with the angular terms involving  $\theta$  and  $\phi$  in Equation (4-10) for  $\nabla^2$ . To find non spherically symmetric solutions, we have to work with the complete equation

$$E\Psi = -\frac{\hbar^2}{2m}\nabla^2\Psi + V(r)\Psi$$

$$\nabla^2\Psi = \frac{1}{r}\frac{\partial^2}{\partial r^2}(r\Psi)$$

$$+\frac{1}{r^2\sin\theta}\frac{\partial}{\partial\theta}\left(\sin\theta\frac{\partial\Psi}{\partial\theta}\right)$$

$$+\frac{1}{r^2\sin^2\theta}\frac{\partial^2\Psi}{\partial\phi^2}$$
(99)

Differential equations involving  $\nabla^2$  in spherical coordinates have been studied for a long time and standard procedures have been carefully worked out to handle the angular dependence of the solutions of these equations. As long as the equation has no other angular terms except those that appear in  $\nabla^2$ , then the solutions are of the form

$$f(\mathbf{r}, \theta, \phi) = R_{n\ell m}(\mathbf{r})Y_{\ell m}(\theta, \phi)$$
(100)

where  $R_{n\ell m}(r)$  are functions that depend only on the variable (r), and the  $Y_{\ell m}(\theta, \phi)$  are functions only of the angles  $\theta$  and  $\phi$ . The subscripts n,  $\ell$  and m can take on only integer values.

When we are dealing with Schrödinger's equation, the solutions are of the form

$$\Psi(\mathbf{r}, \theta, \phi) = \Psi_{\mathbf{n}\ell\mathbf{m}}(\mathbf{r}) \mathbf{Y}_{\ell\mathbf{m}}(\theta, \phi) \tag{101}$$

where each different allowed integer value of the subscripts n, l, and m corresponds to a different allowed standing wave pattern for the electron.

The functions  $Y_{\ell m}(\theta, \phi)$ , which are called *spherical harmonics*, start off quite simply for small  $\ell$ , m, n, but become more complex as  $\ell$  and m increase. The simplest are

$$Y_{0,0}(\theta,\phi) = 1 \qquad (no \ angular \ dependence)$$
$$Y_{1,0} = \cos \theta$$
$$Y_{1,1} = \frac{-1}{\sqrt{2}} \sin \theta \ e^{i\phi}$$
$$Y_{1,-1} = \frac{1}{\sqrt{2}} \sin \theta \ e^{-i\phi} \qquad (102)$$

Since  $Y_{0,0}$  has no angular dependence, all solutions of the form

$$\psi_{n,0,0} = \psi_n(r) Y_{0,0} = \psi_n(r) \tag{103}$$

are the spherically symmetric solutions we have already been studying. We calculated  $\psi_1(r)$  and had you calculate  $\psi_2(r)$ , which corresponds to the values n = 1 and n = 2 respectively.

When we worked out the solution  $\psi_1(r)$  we found that it represented an electron in the lowest, n = 1, energy level. You were to show that  $\psi_2(r)$  represented an electron in the second, n = 2, energy level. We can see that for the symmetric solutions, the integer subscript n is the energy quantum number for the electron.

It turns out that the integer subscripts  $\ell$  and m define the amount of angular momentum the electron has in a particular wave pattern. When  $\ell = 0$ , m = 0, the electron has no angular momentum. Thus the symmetric solutions represent an electron with no angular momentum.

The quantum number  $\ell$  is related to the total orbital angular momentum of the electron, and m is proportional to the z component  $L_z$  of orbital angular momentum. Explicitly

$$L_z = m\hbar \tag{104}$$

The fact that the numbers  $\ell$ , m and n have to have integer values is simply a consequence that for any confined wave, there is an explicit set of allowed standing wave patterns. The electron in the hydrogen atom is confined by the Coulomb force of the proton. When you work out the mathematics to handle  $\nabla^2$  in spherical coordinates, you find that the allowed standing wave patterns can be identified by the integers  $\ell$ , m and n.

There are certain rules for the possible values of  $\ell$ , m and n. When n = 1, there is only one solution which we found. It corresponds to  $\ell = m = 0$ . For n = 2, the possible solutions are:

n	l	m	
2	0	0	possible values of $\ell$ and $m$ for $n = 2$
2	1	0	
2	1	1	
2	1	-1	

In general, n ranges from 1 to infinity,  $\ell$  can have values from 0 up to n - 1, and m can range in integer steps from  $+\ell$  down to  $-\ell$ . These are the rules that define the possible standing wave patterns of the electron in hydrogen.

# Calculus 2000-Chapter 7 Divergence

In the Physics text we pointed out that a vector field was uniquely determined by formulas for the surface integral and the line integral. As we have mentioned several times, that is why there are four Maxwell equations, since we need equations for the surface and line integral of both the electric and magnetic fields. The divergence and curl are the surface and line integrals shrunk down to an infinitesimal or differential scale. We will discuss divergence in this chapter and curl in the next.

## THE DIVERGENCE

As we mentioned, the divergence is a surface integral shrunk down to an infinitesimal or differential scale. To see how this shrinking takes place, we will start with the concept of the surface integral as expressed by Gauss' law and see how we can apply it on a very small scale.

We begin with Equation (29-5) of the Physics text

$$\int_{\substack{\text{closed}\\\text{surface}}} \vec{E} \cdot d\vec{A} = \frac{Q_{\text{in}}}{\varepsilon_0}$$
(29-5)

Equation (29-4) says that for any closed surface, the integral of  $\vec{E} \cdot d\vec{A}$  over the surface is equal to  $1/\epsilon_0$  times the total charge  $Q_{in}$  inside the volume bounded by the surface.

The interpretation we gave to this equation was to call  $\vec{E} \cdot d\vec{A}$  the *flux* of the field  $\vec{E}$  out through the area element  $d\vec{A}$ . The integral over the closed surface is the total flux flowing out through the surface. We said that this net flux out was created by the electric charge inside. By calculating the flux of  $\vec{E}$  out through a spherical surface centered on a point charge, we found that the amount of flux created by a charge Q was  $Q/\epsilon_0$ .

The fact that Equation (29-4) applies to a surface of arbitrary shape follows from the fact that the electric field of a point charge is mathematically similar to the velocity field of a point source in an incompressible fluid like water. We described a point source of a velocity field as some sort of "magic" device that created water molecules. The physical content of Gauss' law applied to water was that the total flux of water out through any closed surface had to be equal to the rate at which water molecules were being created inside.

Of course for a real situation there are no "magic" sources creating water molecules, with the result that there is no net flux of water out through any closed surface, and the velocity field of water obeys the equation

$$\int_{\substack{\text{closed} \\ \text{surface}}} \vec{v} \cdot d\vec{A} = 0 \tag{1}$$

Equation (1) is the condition that the velocity field is a purely solenoidal field like the magnetic field.

Back to Gauss' law, Equation (29-5). Before we shrink the law to an infinitesmal scale, we would like to change the right hand side, expressing the total charge  $Q_{in}$  in terms of the charge density  $\rho(x,y,z)$  that is within the volume bounded by the closed surface.

We do this by considering a small volume element  $\Delta V_i = (\Delta x \Delta y \Delta z)_i$ . If the charge density at point (i) is  $\rho(x_i, y_i, z_i)$  then the amount of charge  $\Delta Q_i$  at  $\Delta V_i$  is

$$\Delta Q_{i} = \rho(x_{i}, y_{i}, z_{i}) \Delta V_{i} \qquad \Delta z \qquad \Delta y \qquad (2)$$

Adding up all the  $\Delta Q_i$  that reside inside the surface gives us

$$Q_{in} = \sum_{i} \Delta Q_{i} = \sum_{i} \rho_{i} \Delta V_{i}$$
  
= 
$$\sum_{i} \rho(x_{i}, y_{i}, z_{i}) \Delta x_{i} \Delta y_{i} \Delta z_{i}$$
(3)

Taking the limit as the  $\Delta x$ ,  $\Delta y$  and  $\Delta z$  go to zero gives us the integral

$$Q_{in} = \int_{\substack{\text{volume} \\ \text{bounded by} \\ \text{closed surface}}} \rho(x,y,z) \, dx \, dy \, dz$$
(4)

To shorten the notation, let V be the volume bounded by the closed surface S, and introduce the notation

$$d^{3}V \equiv dx dy dz \tag{5}$$

Then Equation (4) can be written

$$Q_{in} = \int_{V} \rho(x, y, z) d^{3}V$$
(6)

Using Equation (6) in Gauss' law (29-5) gives us

$$\int_{S} \vec{E} \cdot d\vec{A} = \frac{1}{\varepsilon_0} \int_{V} \rho(x, y, z) d^3 V$$
(7)

Equation (7) is a more general integral form of Gauss' law, relating the surface integral of  $\vec{E}$  over a closed surface S to the volume integral of  $\rho$  over the volume bounded by S. It is Equation (7) that we would now like to shrink down to an infinitesmal scale.

We know how to go to the small scale version of the volume integral of  $\rho$ , just undo the steps (2) through (6) that we used to derive the volume integral. In particular we will focus our attention on one small volume element  $\Delta V_i = \Delta x_i \Delta y_i \Delta z_i$  and apply Gauss' law to this volume

$$\int_{\substack{\text{surface} \\ \text{bounding} \\ \Delta V_i}} \vec{E} \cdot d\vec{A} = \frac{\Delta Q_i}{\epsilon_0} = \frac{1}{\epsilon_0} \rho(x_i, y_i, z_i) \Delta V_i$$
(8)

It is clear how we got the total charge  $Q_{in}$  when we added up all the  $\Delta Q_i$  inside the volume V. But how do we handle the surface integral of  $\vec{E}$ ? How do we interpret adding a bunch of surface integrals over the small volume elements  $\Delta V_i$  to get the surface integral over the entire surface S?

The way to picture it is to remember that the surface integral over the surface of  $\Delta V_i$  is equal to the flux of  $\vec{E}$  created inside  $\Delta V_i$ . From this point of view, the total flux flowing out through the surface of the entire volume will be the sum of the fluxes created within each volume element. To calculate this sum, we first have to calculate the flux flowing out of the volume element  $\Delta V_i$ .

In Figure (1), we show the volume element  $\Delta V_i$ located at  $(x_i, y_i, z_i)$ , with sides  $\Delta x$ ,  $\Delta y$  and  $\Delta z$ . Flowing through this volume element is the electric field  $\vec{E}(x, y, z)$ .

Also in Figure (1) we have drawn the surface area vectors  $\overrightarrow{\Delta A}_1$ , and  $\overrightarrow{\Delta A}_2$  for the left and right vertical faces. Recall that for a surface integral, the area vector  $\overrightarrow{\Delta A}$  or  $d\overrightarrow{A}$  is perpendicular to the surface,

**Figure 1** The volume element  $\Delta V_i$ .

pointing out of the surface. Thus  $\Delta \dot{A}_2$  is x directed with a magnitude equal to the area  $\Delta y \Delta z$  of that side, while  $\overrightarrow{\Delta A}_1$  points in the -x direction and has the same magnitude.

We can formally write

$$\overrightarrow{\Delta A}_1 = -\hat{x}\Delta y\Delta z \; ; \; \overrightarrow{\Delta A}_2 = \hat{x}\Delta y\Delta z \qquad (9)$$

where  $\hat{x}$  is the unit vector in the x direction. Similar formulas hold for the area vectors for the other four faces of  $\Delta V$ . For example, on the top face we have  $\overrightarrow{\Delta A}_3 = \hat{z} \Delta x \Delta y$ .

To calculate the total flux of  $\vec{E}$  out of  $\Delta V$ , we have to calculate the flux out through each of the six faces. For the two x oriented areas  $\vec{\Delta A_1}$ , and  $\vec{\Delta A_2}$ , only the x component of  $\vec{E}$  will contribute to the dot products  $\vec{E} \cdot \vec{\Delta A}$ . Let  $\vec{E}_x(x,y,z)$  be the average value of  $E_x$  at face 1, and  $\vec{E}_x(x + \Delta x, y, z)$  be the average value of  $E_x$  at face 2, which is a distance  $\Delta x$ down the x axis from face 1. The flux out of face 2 will be

$$flux outof face 2 = \overline{E}_{x}(x + \Delta x, y, z)\Delta A_{2}$$
$$= \overline{E}_{x}(x + \Delta x, y, z)\Delta y\Delta z$$
(10)

At face 1, where  $\overrightarrow{\Delta A}_1 = -\hat{x}\Delta y\Delta z$ , the dot product  $\vec{E}\cdot\vec{\Delta A}$  can be written

$$\vec{E} \cdot \vec{\Delta A}_{1} = (\hat{x} \vec{E}_{x} + \hat{y} \vec{E}_{y} + \hat{z} \vec{E}_{z}) \cdot (-\hat{x} \Delta y \Delta z)$$

$$= -\vec{E}_{x}(x, y, z) \Delta y \Delta z$$
(11)

where  $\hat{\mathbf{x}} \cdot \hat{\mathbf{x}} = 1$ ,  $\hat{\mathbf{y}} \cdot \hat{\mathbf{x}} = \hat{\mathbf{z}} \cdot \hat{\mathbf{x}} = 0$ . We wrote the full dot product in Equation (11) so that you could see explicitly where the minus sign came from.

Combining Equations (10) and (11) for the total flux out of the two x directed faces of  $\Delta V$ , we get

$$\int_{\text{faces}}^{\text{flux out}} = \left[ \overline{E}_{x}(x + \Delta x, y, z) - \overline{E}_{x}(xy, z) \right] \Delta y \Delta z$$

$$\int_{\text{faces}}^{\text{flux out}} \Delta y \Delta z$$

$$(12)$$



If we multiply Equation (12) by  $\Delta x / \Delta x = 1$  we get

$$\begin{aligned} & \underset{\text{directed}}{\underset{\text{faces of }\Delta V}{\text{of }x}} = \left[ \frac{\overline{E}_{x}(x + \Delta x, y, z) - \overline{E}_{x}(x, y, z)}{\Delta x} \right] \Delta x \Delta y \Delta z \end{aligned}$$

$$(13)$$

At this point,  $\overline{E}_x(x + \Delta x, y, z)$  and  $\overline{E}_x(x, y, z)$  are the average values of  $\overline{E}_x$ , averaged over the x directed faces at  $x + \Delta x$  and x respectively, while the functions without averaging, namely  $E_x(x + \Delta x, y, z)$  and  $E_x(x, y, z)$  are just the values of  $E_x$  at the lower front corners of the x oriented faces as shown in Figure (2). Any difference between the average values of  $\overline{E}_x$  and the corner values  $E_x$  will be due to y and z variations of  $E_x$  over the area  $\Delta y \Delta z$ .

In Equation (13) we see that the change of  $E_x$ , as we move in the x direction, is going to become very important. It should be clear that we are going to get a partial derivative of  $E_x$  with respect to x. What we are going to do now is say that variations of  $E_x$  in the x direction are important but variations of  $E_x$  in the y and z direction are not, and as a result we can replace the average values of  $\overline{E}_x$  with the corner values  $E_x$ .

The above paragraph was intended to sound like a questionable procedure. If we do it, Equation (13) immediately simplifies, as we will see shortly. But how do we justify such a step? The answer, which we work out in detail in the appendix to this chapter, is that when we take the limit as  $\Delta V$  goes to zero, contributions due to y and z variations of  $E_x$  go to zero faster than the contribution from the x variation. Neglecting the y and z variations turns out to be similar to neglecting  $\alpha^2$  terms compared to  $\alpha$  terms in an expansion of  $(1 + \alpha)^n$  when  $\alpha$  is a small number.



**Figure 2** *Electric field at the lower front corners.* 

We put this discussion in the appendix because it takes some effort which distracts from our goal of reducing Gauss' law to a differential equation. However it is important to know how to figure out when certain terms or dependencies can be neglected when we take calculus limits. Thus the appendix should not be skipped.

Assuming that we can replace  $\overline{E}_x$  by  $E_x$  in Equation (13), noting that  $\Delta x \Delta y \Delta z = \Delta V$ , and taking the limit as  $\Delta x$  goes to zero gives us

flux out of  
x directed  
faces of 
$$\Delta V = \frac{\text{limit}}{\Delta x \rightarrow 0} \left[ \frac{E_x(x + \Delta x, y, z) - E_x(x, y, z)}{\Delta x} \right] \Delta V$$
(14)

The limit is clearly the partial derivative  $\partial E_x(x,y,z)/\partial x$  and we get

flux out of  
x directed faces of 
$$\Delta V = \frac{\partial E_x(x,y,z)}{\partial x} \Delta V$$
 (15a)

Similar equations should apply to the y and z faces, giving us

flux out of  
y directed = 
$$\frac{\partial E_y(x,y,z)}{\partial y} \Delta V$$
 (15b)  
faces of  $\Delta V$ 

flux out of  
z directed faces of 
$$\Delta V = \frac{\partial E_z(x,y,z)}{\partial z} \Delta V$$
 (15c)

### Exercise 1

Draw the appropriate sketches and reproduce the arguments needed to derive Equation (15b) or (15c).

When we add up the flux out of all six faces, we get the total flux out of  $\Delta V$ 

total flux  
out of 
$$\Delta V = \left[ \frac{\partial E_x}{\partial x} + \frac{\partial E_y}{\partial y} + \frac{\partial E_z}{\partial z} \right] \Delta V$$
 (16)

You should spot immediately that the notation in Equation (16) can be simplified by introducing the partial derivative operator

$$\vec{\nabla} = \left(\hat{x}\frac{\partial}{\partial x} + \hat{y}\frac{\partial}{\partial y} + \hat{z}\frac{\partial}{\partial z}\right)$$
  
$$\equiv (\hat{x}\nabla_{x} + \hat{y}\nabla_{y} + \hat{z}\nabla_{z})$$
(17)

From the definition of the vector dot product we have

$$\vec{\nabla} \cdot \vec{E} = \left(\hat{x}\frac{\partial}{\partial x} + \hat{y}\frac{\partial}{\partial y} + \hat{z}\frac{\partial}{\partial z}\right) \cdot (\hat{x}E_x + \hat{y}E_y + \hat{z}E_z)$$
$$= \frac{\partial E_x}{\partial x} + \frac{\partial E_y}{\partial y} + \frac{\partial E_z}{\partial z}$$
(18)

where we used  $\hat{\mathbf{x}} \cdot \hat{\mathbf{x}} = 1$ ,  $\hat{\mathbf{x}} \cdot \hat{\mathbf{y}} = 0$ , etc., and noted that the unit vectors are constants that can be taken outside the derivative. For example,

$$\frac{\partial}{\partial x}(\hat{x} E_x) = \hat{x} \frac{\partial E_x}{\partial x}$$
(18a)

Using the notation of Equation (18), we get for the total flux out of  $\Delta V$ 

$$\begin{array}{l} \text{total flux}\\ \text{out of } \Delta V \end{array} = (\vec{\nabla} \cdot \vec{E}) \Delta V \tag{19} \end{array}$$

Equation (19) applies to each  $\Delta V_i$  at each point  $(x_i, y_i, z_i)$  within any volume V bounded by a closed surface S. The total flux out through the surface S, which is the surface integral of  $\vec{E}$ , will be equal to the sum of all the flux created inside in all the  $\Delta V_i$ . Thus we get

$$\int_{\substack{\text{surface}\\\text{bounding V}}} \vec{E} \cdot d\vec{A} = \sum_{i} (\vec{\nabla} \cdot \vec{E}) \Delta V_{i}$$
(20)

As we take the limit at  $\Delta V_i$  goes to zero size, the sum becomes an integral, and we end up with

$$\int_{\substack{\text{closed surface} \\ \text{bounding} \\ \text{volume V}}} \vec{E} \cdot d\vec{A} = \int_{V} \vec{\nabla} \cdot \vec{E} d^{3}V \qquad \begin{array}{c} \text{divergence} \\ \text{theorem} \\ \text{theorem} \end{array} (21)$$

where we are using the notation of Equation (5) that  $d^{3}V \equiv dx dy dz$ .

Equation (21) is known as the *divergence theorem*, and the quantity  $\vec{\nabla} \cdot \vec{E}$  is known as the *divergence* of the vector field  $\vec{E}$ . We saw the same operator  $\vec{\nabla}$  in the Chapter 3 when it acted on a scalar field f(x,y,z). Then we had what was called a *gradient* 

$$\vec{\nabla} f \qquad \begin{array}{l} gradient \ of \\ a \ scalar \ field \end{array}$$

$$\vec{\nabla} \cdot \vec{E} \qquad \begin{array}{l} divergence \ of \\ a \ vector \ field \end{array}$$
(22)

You can see that  $\vec{\nabla}$  operating on a scalar field f(x,y,z) creates a vector field  $\vec{\nabla} f$ . In contrast, the dot product of  $\vec{\nabla}$  with a vector field  $\vec{E}$  creates a scalar field  $\vec{\nabla} \cdot \vec{E}$  that has a value at every point in space but does not point anywhere.

Equation (21), the divergence theorem, is an extremely useful result for it allows us to go back and forth between a surface integral and a volume integral. In Equation (7) reproduced here,

$$\int_{S} \vec{E} \cdot d\vec{A} = \frac{1}{\epsilon_0} \int_{V} \rho(x,y,z) d^3 V \qquad (7) \text{ repeated}$$

we had a mixed bag with a surface integral over a closed surface on the left and a volume integral over the enclosed volume V on the right. Back then, there was not much more we could do with that equation.

But now we can replace the surface integral of  $\vec{E}$  with a volume integral of  $\vec{\nabla} \cdot \vec{E}$  to get

$$\int_{S} \vec{E} \cdot d\vec{A} = \frac{1}{\epsilon_{0}} \int_{V} \rho(x,y,z) d^{3}V \quad (7) \text{ repeated}$$
$$\int_{V} \vec{\nabla} \cdot \vec{E} d^{3}V = \frac{1}{\epsilon_{0}} \int_{V} \rho(x,y,z) d^{3}V \quad (23)$$

Since we are integrating over the same volume V for both integrals, we can write (23) as

$$\int_{V} \left[ \vec{\nabla} \cdot \vec{E}(x,y,z) - \frac{\rho(x,y,z)}{\epsilon_0} \right] d^3V = 0 \qquad (24)$$

The next argument is one often used in physics. Since the integral in Equation (24) has to be zero for any volume V we choose, the only way that can happen is if the integrand, the stuff in the square brackets, is zero. This gives us the differential equation

$$\vec{\nabla} \cdot \mathbf{E}(\mathbf{x}, \mathbf{y}, \mathbf{z}) = \frac{\rho(\mathbf{x}, \mathbf{y}, \mathbf{z})}{\varepsilon_0} \begin{bmatrix} Gauss' \\ law in \\ differential \\ form \end{bmatrix} (25)$$

Equation (25) is the differential equation representing Gauss' law. When Maxwell's equations are written as differential equations, this will be one of the four.

## **Exercise 2**

Another of Maxwell's equations in integral form is

$$\int_{\substack{\text{closed}\\\text{surface}}} \vec{B} \cdot d\vec{A} = 0$$

What is the corresponding differential equation?

## **Electric Field of a Point Charge**

Until now, in both the Physics and Calculus texts, when we obtained a new differential equation, we illustrated its use with explicit examples. This time we do not yet have a good example for our new Equation (25)  $\vec{\nabla} \cdot \vec{E} = \rho/\epsilon_0$ . This is the differential form of Gauss' law, and our best example for the use of Gauss' law was in calculating the electric field of a point charge. The problem is that, at the point charge itself, the field  $\vec{E}$  and its partial derivatives are infinite and the assumptions we made in deriving Equation (25) do not apply.

When we are dealing with the electric field of a point charge, the field  $\vec{E}$  is well behaved and all partial derivatives are finite, except at the charge. The way we can handle point charges is to use Equation (25)  $\vec{\nabla} \cdot \vec{E} = \rho/\epsilon_0$  everywhere except in a small region around the charge. In that region we revert to the integral form of Gauss' law which allows us to work just outside the point charge and avoid the infinities.

Here is an outline of the way we handle the problem of a point charge. We are working with Equation (25)

$$\vec{\nabla} \cdot \mathbf{E}(\mathbf{x},\mathbf{y},\mathbf{z}) = \frac{\rho(\mathbf{x},\mathbf{y},\mathbf{z})}{\varepsilon_0}$$
 (25) repeated

and everything is going well until we come up to a point charge located at the point  $(x_0,y_0,z_0)$ . In a small region surrounding the point charge, we integrate Equation (25) over the volume, getting

$$\int_{\substack{\text{volume}\\\text{surrounding}\\\text{charge}}} \vec{\nabla} \cdot \vec{E} \, d^3 V = \int_{\substack{\text{volume}\\\text{surrounding}\\\text{charge}}} \frac{\rho}{\epsilon_0} d^3 V \quad (26)$$

The volume integral of the charge density  $\rho$  over the region of the point charge is simply the charge Q itself, thus we can immediately do that volume integral, giving us

$$\int_{\substack{\text{volume} \\ \text{surrounding} \\ \text{charge}}} \vec{\nabla} \cdot \vec{E} \, d^3 V = \frac{Q}{\epsilon_0}$$
(27)

We still have the problem that  $\vec{\nabla} \cdot \vec{E}$  is infinite at the charge itself. But we can avoid this problem by converting the volume integral of  $\vec{\nabla} \cdot \vec{E}$  to a surface integral of  $\vec{E}$  using the divergence theorem, Equation (21)

 $\int_{\substack{\text{volume}\\\text{surrounding}\\\text{charge}}} \vec{\nabla} \cdot \vec{E} \, d^3 V = \int_{\substack{\text{surface}\\\text{enclosing}\\\text{charge}}} \vec{E} \cdot d\vec{A}$ (21) repeated to get

$$\int_{\substack{\text{surface}\\ \text{surrounding}\\ \text{charge}}} \vec{E} \cdot d\vec{A} = \frac{Q}{\epsilon_0}$$
(28)

In Equation (28), which we recognize as the form of Gauss' law we started with in the Physics text, the electric field is evaluated only at the surface surrounding the point charge, and not at the charge itself. Away from the charge, the field is finite and we have no problem with Equation (28).

There is a mathematical problem with the concept of a point charge, where a finite amount of charge is crammed into a region of zero volume, giving us infinite charge densities and infinite fields there. We have just shown how these infinities can be avoided mathematically, at least for Gauss' law, by converting the volume integral of  $\vec{\nabla} \cdot \vec{E}$  at the charge to a surface integral of  $\vec{E}$  out from the charge. Was this just a mathematical exercise, or in physics do we really have to deal with point charges?

The theory of quantum electrodynamics, which describes the interaction of electrons with light (with photons), is the most precisely verified theory in science. It explains, for example, the very smallest relativistic corrections observed in the spectrum of the hydrogen atom. This theory treats the electron as an actual *point particle* with a finite amount of mass and charge confined to a region of zero volume. The trick we just pulled to handle the electric field of a point charge was quite simple compared to the tricks that the inventors of quantum electrodynamics, Feynman, Schwinger, and Tomonaga, had to pull to handle the infinite mass and energy densities they encountered. The remarkable accomplishment was that they succeeded in constructing a theory of point particles, a theory that gave finite and correct, answers.

The question that remains unanswered, is whether the electron is truly a point particle, or does it have some size that is so small that we have not been able to see the structure yet? The important feature of quantum electrodynamics is that it makes testable predictions without any reference to the electron's structure. We get the same predictions whether the electron has no size, or is some structure that is too small to see. Our handling of the electric field of a point charge is your first example of how such a theory can be constructed. By converting to a surface integral surrounding the charge, it makes no difference whether the charge is truly a point, or confined to some region too small to see.

By the way, in the current picture of elementary particles, in what is often called the *standard model*, the true elementary particles are all point particles. These elementary particles are the six electron type particles called *leptons* (they are the electron, the muon, the tau particle, and three kinds of neutrinos) and six kinds of *quarks*. The standard model makes many successful predictions but appears to have one critical flaw. The problem is that no one has yet succeeded in constructing a theory for the interaction of point particles with gravity, the so called *quantum theory of gravity*. Every attempt to do so has thus far led to infinities that could not be gotten rid of by any known mathematical technique.

This failure to develop a quantum theory of gravity in which gravity interacts with point particles, has led to theories such as *string theory* where the elementary particles have a finite, but tiny size. String theory appears to avoid the infinities in the gravitational interaction, but the strings, from which particles are assumed to be made, are predicted to be so small that no way has been found to test whether they actually exist or not. It is interesting that so far our only evidence that elementary particles actually have structure is our failure to construct a theory of gravity.

## THE $\delta$ FUNCTION

When we applied the differential form of Gauss' law  $\vec{\nabla} \cdot \vec{E} = \rho/\epsilon_0$  to the field of a point charge, we avoided the problem of mathematical infinities by integrating the equation over a small volume surrounding the charge. We never did say what the charge density  $\rho(x,y,z)$  was for a point charge Q, because we knew that if we integrated  $\rho(x,y,z)$  over the region of the charge, the answer would be simply Q itself.

In physics we often run into quantities like the charge density of a point charge where the density at the charge looks infinite, but when we integrate the density over the region of the charge, we get a finite, reasonable answer. There is a convenient way to handle such problems by using what is called the *delta* ( $\delta$ ) *function*.

The one dimensional  $\delta$  function is a curve with a unit area under it, but all the area is confined to a region of zero width. We obtain such a curve mathematically through the use of a limiting process.

Consider the curve shown in Figure (3) that is zero everywhere except in the region around the point  $x_0$ . In that region it is a rectangle of width  $\Delta x$  and height  $1/\Delta x$ . The area under this curve is

area under  
rectangle = 
$$(\Delta x) \left( \frac{1}{\Delta x} \right) = 1$$
 (29)  
 $\overrightarrow{Ax} = 1$   
 $\overrightarrow{Ax} =$ 

When we take the limit as  $\Delta x$  goes to zero, we get a one dimensional delta function.

Now take the limit as  $\Delta x \rightarrow 0$ , and we end up with a curve, whose total area remains 1, but whose width goes to zero and height goes to infinity. We will call this curve  $\delta(x_0)$ 

$$\delta(\mathbf{x}_{0}) \equiv \lim_{\Delta \mathbf{x} \to 0} \left( \begin{array}{c} \text{of the curve of width} \\ \Delta \mathbf{x} \text{ and height } 1/\Delta \mathbf{x}, \\ \text{centered at } \mathbf{x}_{0} \end{array} \right)$$
(30)

Even though  $\delta(x_0)$  is infinitely high at the point  $x_0$ , its integral over any region that includes the point  $x_0$  is just the number 1

x greater than 
$$x_0$$
  

$$\int_{x \text{ less than } x_0} \delta(x_0) dx = 1$$
(31)

Actually the only important property of the  $\delta$  function is Equation (31). The curve does not have to be a rectangle, it could be the limit of some smooth curve like that shown in Figure (4). As long as, in the limit that  $\Delta x \rightarrow 0$ , the curve becomes infinitely high, infinitely narrow, and has a unit area under it, it is a  $\delta$  function.

In three dimensions, the  $\delta$  function  $\delta(x_0, y_0, z_0)$  is a quantity that is zero everywhere except at the point  $(x_0, y_0, z_0)$ , but whose integral over that region is 1

$$\int_{\text{any volume}} \delta(x_0, y_0, z_0) dV = 1$$
including the
point (x\_0, y\_0, z\_0)
(32)

An example of such a  $\delta$  function is the function whose value is zero everywhere except within a distance  $\Delta x$  of  $x_0$ ,  $\Delta y$  of  $y_0$ , and  $\Delta z$  of  $z_0$ . In that region the value is  $(1/\Delta x)(1/\Delta y)(1/\Delta z)$ , so that the total volume is 1. Then take the limit as  $\Delta x \rightarrow 0$ ,  $\Delta y \rightarrow 0$ , and  $\Delta z \rightarrow 0$ .



**Figure 4** We have a delta function as long as the area remains 1, and the width goes to zero.

We can now use the  $\delta$  function to describe the charge density of a point charge. If a point charge has a total charge Q and is located at the point  $(x_0,y_0,z_0)$ , then the charge density  $\rho(x,y,z)$  is

$$\rho(\mathbf{x},\mathbf{y},\mathbf{z}) = \mathbf{Q}\delta(\mathbf{x}_0,\mathbf{y}_0,\mathbf{z}_0) \begin{array}{l} \text{charge density} \\ \text{of point charge} \\ \text{at } \mathbf{x}_0,\mathbf{y}_0,\mathbf{z}_0 \end{array} (33)$$

The differential form of Gauss' law applied to this charge density is

$$\vec{\nabla} \cdot \vec{E} = \frac{\rho(x, y, z)}{\varepsilon_0}$$

$$\vec{\nabla} \cdot \vec{E} = \frac{Q}{\varepsilon_0} \delta(x_0, y_0, z_0)$$
(34)

To handle Equation (34), we use our old trick of going back to the integral form by first integrating over a volume that includes the charge

$$\int_{\substack{\text{volume} \\ \text{including} \\ \text{charge}}} \vec{\nabla} \cdot \vec{E} \, dV = \int_{\substack{\text{volume} \\ \text{including point} \\ x_0, y_0, z_0}} \frac{Q}{\epsilon_0} \delta(x_0, y_0, z_0) dV$$
(35)

Since  $Q/\varepsilon_0$  is a constant, it can be taken outside the integral on the right side of Equation (35), giving

$$\frac{Q}{\epsilon_{0}} \int_{\substack{\text{volume} \\ \text{including point } x_{0}, y_{0}, z_{0}}} \delta(x_{0}, y_{0}, z_{0}) dV = \frac{Q}{\epsilon_{0}} \times 1$$
(36)

where we used the fact that the integral of the  $\delta$  function was 1. Now convert the volume integral of  $\vec{\nabla} \cdot \vec{E}$  to a surface integral

$$\int_{\substack{\text{volume}\\\text{including point\\x_0, y_0, z_0}} \vec{\nabla} \cdot \vec{E} \, dV = \int_{\substack{\text{surface}\\\text{surrounding}\\x_0, y_0, z_0}} \vec{E} \cdot d\vec{A}$$
(37)

Using (36) and (37) gives

$$\int_{\substack{\text{closed surface}\\\text{including }Q}} \vec{E} \cdot d\vec{A} = \frac{Q}{\epsilon_0}$$

which is our integral form of Gauss' law.

From this example, you can see that the  $\delta$  function allows us to write an explicit formula for the charge density of a point charge, and you can see that the only things we have to know about a  $\delta$  function is that  $\delta(x_0, y_0, z_0)$  is zero except at  $(x_0, y_0, z_0)$  and that its volume integral around that point is 1. As you go farther in physics, you will encounter the  $\delta$ function more and more often. It is rather nice in that there is no function easier to integrate.

## **Exercise 3**

Explain why the following mathematical relationship is true for any continuous function f(x,y,z)

$$\int f(x,y,z)\delta(x_0,y_0,z_0)d^3V = f(x_0,y_0,z_0)$$
any volume  
including the  
point (x\_0,y\_0,z\_0)
(38)

## **DIVERGENCE FREE FIELDS**

It may seem a bit discouraging that we did all this work to derive the differential form of Gauss' law  $\vec{\nabla} \cdot \vec{E} = \rho/\epsilon_0$ , and then end up, when we want to actually solve a problem, going back to the integral form of the equation. At this point, that is about all we can do to solve for explicit field patterns  $\vec{E}$ . However, the differential form begins to tell us about some general features of a vector field as we shall now see. With a lot more practice with the differential form of the field equations, and perhaps a computer thrown in, one can begin to solve for complex field shapes. In this text we will focus on what we can learn about general features and leave the solution of complex field shapes to a later course.

To see what we can learn about general features of a field, suppose that we have a velocity field  $\vec{v}(x,y,z)$ , whose divergence is zero, i.e., it obeys the equation

$$\vec{\nabla} \cdot \vec{v}(\mathbf{x}, \mathbf{y}, \mathbf{z}) = 0 \tag{39}$$

We say that such a field is *divergence free*. What can we say about the properties of such a field?

To answer that question, we will again go back to the integral form, by integrating Equation (1) over some volume V to get

$$\int_{\text{volume V}} \vec{\nabla} \cdot \vec{v} \, d^3 V = 0 \tag{40}$$

Now use the divergence theorem to convert this volume integral to a surface integral, giving

$$\int_{\substack{\text{closed}\\\text{surface}}} \vec{v} \cdot d\vec{A} = 0 \tag{41}$$

Equation (41) is our old equation for a vector field that has no sources or sinks. It is the equation for an incompressible, constant density fluid, a real one like water where water molecules are not being created or destroyed. Thus the condition that a vector field be divergence free, i.e.,  $\vec{\nabla} \cdot \vec{v} = 0$  or  $\vec{\nabla} \cdot \vec{E} = 0$  or  $\vec{\nabla} \cdot \vec{B} = 0$ , is that the field behaves like the velocity field of an incompressible fluid. What kind of solutions are possible for a divergence free field? What are the solutions to the equation  $\vec{\nabla} \cdot \vec{v} = 0$ ?

The answer is at least as complex as the behavior of water. You have seen water flow smoothly in a lazy river. That is called *laminar flow*. Such laminar flow is one solution to  $\vec{\nabla} \cdot \vec{v} = 0$ . But in a fast flowing stream there can be complex eddies called *turbulence*. Turbulent flow is also a solution to the equation  $\vec{\nabla} \cdot \vec{v} = 0$ .

You can now see that the equation  $\vec{\nabla} \cdot \vec{v} = 0$  puts a restriction on the field  $\vec{v}$ , but still allows an enormous range of solutions. Because of your familiarity with the flow of water you have some insight into what these solutions can be.

## APPENDIX — DERIVATION OF FLUX EQUATION (14)

Earlier in the chapter we had the following formula for the flux out of the x directed faces of the small cube  $\Delta V = \Delta x \Delta y \Delta z$ 

flux out  
of x directed = 
$$\left[\frac{\overline{E}_{x}(x+\Delta x, y, z) - \overline{E}_{x}(x, y, z)}{\Delta x}\right] \Delta x \Delta y \Delta z$$
(12)

(13) repeated where  $\overline{E}_x(x+\Delta x,y,z)$  and  $\overline{E}_x(x,y,z)$  were the average values of  $E_x$  on the two x directed faces of the cube.

In Equation (14) we replaced the average values  $\overline{E}_x$  by the values  $E_x(x+\Delta x,y,z)$  and  $E_x(x,y,z)$  at the lower front corners as shown in Figure (2), repeated here giving

$$\begin{aligned} & \text{flux out} \\ & \text{of x directed} = \underset{\Delta x \to 0}{\text{limit}} \left[ \frac{E_x(x + \Delta x, y, z) - E_x(x, y, z)}{\Delta x} \right] \Delta V \end{aligned}$$

(14) repeated

What we are doing is in going from Equation (13) to (14) is to neglect the y and z dependence of  $E_x$  while developing an equation for the x dependence. This step needs justification.

To see what effect the y and z dependence has, let us start by approximating the average value of  $E_x$  over the entire x faces by the average of the top and bottom values of the front side of  $\Delta A_X$ , i.e., the average of  $E_x$  at points (1) and (3) on the left and points (2) and (4) on the right as shown in Figure (5).

This is a rather crude approximation for the average over the face, but begins to show us what the effect of the y and z dependence of  $E_x$  is.



**Figure 2 (repeated)** *Electric field at the lower front corners.* 

To evaluate  $E_x$  at  $(x, y, z + \Delta z)$ , up at point (3), we can use a Taylor series expansion. So far we have discussed a Taylor series expansion only of a function of a single variable f(x). The expansion was, from Equation (2-44 of Calculus Chapter 2)

$$f(x-x_0) = f(x_0) + \frac{\partial f}{\partial x}(x-x_0) + \frac{1}{2!}\frac{\partial^2 f}{\partial x^2}(x-x_0)^2 + \cdots$$
(2-44) repeated

which is good for small steps  $(x-x_0)$ .

What we are doing when we go from point (1) to point (3) in Figure (2), is keeping the values of x and y constant, and looking at the change in  $E_x$  as we vary z. Thus in going up, we have a function  $E_x(z)$ that is only a function of z, and we can use our old Taylor series expansion to get

$$E_{x}(x,y,z+\Delta z) = E_{x}(x,y,z) + \frac{\partial E_{x}(x,y,z)}{\partial z} (\Delta z) + \frac{1}{2} \frac{\partial^{2} E_{x}(x,y,z)}{\partial z^{2}} (\Delta z)^{2}$$
(42)

where  $\Delta z$  is analogous to the step  $(x-x_0)$  in the Taylor series formula.

Because we are eventually going to take the limit as  $\Delta z$  goes to zero, we will be able to neglect terms of order  $(\Delta z)^2$  compared to  $\Delta z$ . Because of that, it is sufficient to write

$$E_{x}(x,y,z+\Delta z) = E_{x}(x,y,z) + \frac{\partial E_{x}}{\partial z} \Delta z \qquad (42a)$$
  
+ terms of order  $\Delta z^{2}$ 



**Figure 5** *Electric field at four positions.* 

When we take the average of  $E_x$  at points (1) and (3), a result we will call  $\overline{E}_x(x)_{1,3}$ , we get

$$\begin{split} \overline{E}_{x}(x)_{1,3} &= \frac{E_{x}(x,y,z) + E_{x}(x,y,z+\Delta z)}{2} \\ &= E_{x}(x,y,z) + \frac{1}{2} \frac{\partial E_{x}(x,y,z)}{\partial z} \Delta z + O(\Delta z^{2}) \\ \end{split}$$

where  $O(\Delta z^2)$  means terms of order  $(\Delta z^2)$ .

A similar argument gives the average  $\overline{E}_x(x+\Delta x)_{2,4}$  at points (2) and (4)

$$\overline{E}_{x}(x+\Delta x)_{2,4} = \frac{E_{x}(x+\Delta x, y, z) + E_{x}(x+\Delta x, y, z+\Delta z)}{2}$$
$$= E_{x}(x+\Delta x, y, z) + \frac{1}{2} \frac{\partial E_{x}(x+\Delta x, y, z)}{\partial z} \Delta z + O(\Delta z^{2})$$
(44)

Using our 2 point averages in Equation (13) for the flux out of  $\Delta V$  gives us

flux out of  
x directed face  
of 
$$\Delta V$$
 for 2  
point average =  $\left[\frac{\overline{E}_{x}(x+\Delta x)_{2,4} - \overline{E}_{x}(x)_{1,3}}{\Delta x}\right]\Delta V$ 

$$= \frac{\Delta V}{\Delta x} \left[ E_x(x + \Delta x, y, z) + \frac{1}{2} \frac{\partial E_x(x + \Delta x, y, z)}{\partial z} \Delta z - E_x(x, y, z) - \frac{1}{2} \frac{\partial E_x(x, y, z)}{\partial z} \Delta z + O(\Delta z^2) \right]$$

$$= \left[\frac{E_{x}(x+\Delta x,y,z) - E_{x}(x,y,z)}{\Delta x}\right] \Delta V$$
$$+ \frac{1}{2} \left[\frac{\frac{\partial E_{x}(x+\Delta x,y,z)}{\partial z} - \frac{\partial E_{x}(x,y,z)}{\partial z}}{\Delta x}\right] \Delta z \Delta V \quad (45)$$
$$+ O(\Delta z^{2}) \Delta V$$

When we go to the limit that  $\Delta x$  goes to zero, we see that we get the partial derivatives

$$\lim_{\Delta x \to 0} \left[ \frac{E_x(x + \Delta x, y, z) - E_x(x, y, z)}{\Delta x} \right] = \frac{\partial E_x(x, y, z)}{\partial x}$$
(46)

$$\lim_{\Delta x \to 0} \left[ \frac{\frac{\partial E_x(x + \Delta x, y, z)}{\partial z} - \frac{\partial E_x(x, y, z)}{\partial z}}{\Delta x} \right] = \frac{\partial^2 E_x(x, y, z)}{\partial x \partial z}$$
(47)

Thus Equation (45) is taking on the form

We see that corrections due to the z dependence of  $E_x$  are of magnitude  $\Delta z$  times the partial second derivative  $\partial^2 E_x / \partial x \partial z$ . As long as all derivatives of  $E_x$  are bounded, stay finite as we take the limit as  $\Delta x$ ,  $\Delta y$ , and  $\Delta z$  go to zero, then the  $\Delta z$  term in Equation (48) becomes negligently small, which means that in the limit we can neglect the z dependence of  $E_x$ , at least in this two point approximation.

Our 2 point approximation to the average of  $E_x$  can be improved by using more points. If we included the back points at  $(y+\Delta y)$ , we would add terms to Equation (48) of the form

$$\frac{\partial^2 E_x}{\partial x \partial y} \Delta y + O(\Delta z^2)$$
(49)

terms which would go to zero in the limit  $\Delta y \rightarrow 0$ . All points we add in to the average will give terms proportional to  $\Delta x$  or  $\Delta y$  or some combination, and all these terms will go to zero when we take the limit as  $\Delta x$ ,  $\Delta y$ , and  $\Delta z$  goes to zero. Thus, it is an exact result that, in the limit that  $\Delta v \rightarrow 0$ , only the x dependence of  $E_x$  has to be taken into account, provided all derivatives of  $E_x$  are finite.

# Calculus 2000-Chapter 8 Curl

## **ABOUT THE CURL**

In the Physics text, we saw that a vector field was uniquely determined by formulas for the surface integral and the line integral. In the last chapter, we saw that the divergence, such as  $\vec{\nabla} \cdot \vec{E}$ , represented the surface integral shrunk down to an infinitesimal scale. In this chapter, we study the curl, which is the line integral shrunk down to an infinitesimal scale. *Here our emphasis will be on the application of the* curl to electric and magnetic fields. In the final chapters of this text, Chapters 12 and 13, we develop an intuitive picture of the curl applied to the velocity field of fluids such as water and superfluid helium. The curl of the velocity field is called vorticity, a concept that plays a fundamental role in understanding such phenomena as quantum vortices and turbulence.

## INTRODUCTION TO THE CURL

The partial derivative operator

$$\vec{\nabla} = \hat{x}\frac{\partial}{\partial x} + \hat{y}\frac{\partial}{\partial y} + \hat{z}\frac{\partial}{\partial z}$$

has now appeared in our formulas for the *gradient* of a scalar field f(x,y,z)

$$\vec{\nabla}f(x,y,z) = \hat{x}\frac{\partial f}{\partial x} + \hat{y}\frac{\partial f}{\partial y} + \hat{z}\frac{\partial f}{\partial z}$$
(1)

in the *divergence* of a vector field  $\vec{E}(x,y,z)$ 

$$\vec{\nabla} \cdot \vec{E} = \frac{\partial E_x}{\partial x} + \frac{\partial E_y}{\partial y} + \frac{\partial E_z}{\partial z}$$
(2)

and in the Laplacian

$$\vec{\nabla} \cdot \vec{\nabla} f = \nabla^2 f = \frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2} + \frac{\partial^2 f}{\partial z^2}$$
(3)

While  $\vec{\nabla}$  is an operator in the sense that it only has a value when operating on some field, we see that it acts very much like a vector. This suggests that we may encounter other vector like operations involving  $\vec{\nabla}$ .

In our discussion of vectors in Chapter 2 of the Physics text, we saw that there were two kinds of vector products, the scalar or dot product

$$\mathbf{C} = \vec{\mathbf{A}} \cdot \vec{\mathbf{B}} = (\mathbf{A}_{\mathbf{x}} \mathbf{B}_{\mathbf{x}} + \mathbf{A}_{\mathbf{y}} \mathbf{B}_{\mathbf{y}} + \mathbf{A}_{\mathbf{z}} \mathbf{B}_{\mathbf{z}}) \frac{scalar}{product}$$
(4)

and the vector cross product

$$\vec{C} = \vec{A} \times \vec{B}$$
 vector cross product (5)

where the formulas for the components of  $\vec{C}$  were

$$C_{x} = A_{y}B_{z} - A_{z}B_{y}$$

$$C_{y} = A_{z}B_{x} - A_{x}B_{z}$$

$$C_{z} = A_{x}B_{y} - A_{y}B_{x}$$
(6)

We saw that the vector  $\vec{C} = \vec{A} \times \vec{B}$  was oriented perpendicular to the plane of the vectors  $\vec{A}$  and  $\vec{B}$ , the choice of which direction being given by the right hand rule as shown in Figure (1). The magnitude was  $C = AB \sin \theta$  which is maximum when  $\vec{A}$ and  $\vec{B}$  are perpendicular and zero when parallel.

The vector cross product seems like a rather peculiar mathematical construct, but it plays an important role in physics, particularly in describing rotational motion. You will recall that the angular analogy to Newton's second law was

$$\vec{\tau} = \frac{d\vec{L}}{dt} \tag{7}$$

where the torque,  $\vec{\tau} = \vec{r} \times \vec{F}$ , is what we called the angular force, and  $\vec{L} = \vec{r} \times \vec{p}$  is the angular momentum. Despite the appearance of two cross products in Equation (7), the equation led to a very successful prediction of the motion of a gyroscope at the end of Chapter 12 in the Physics text (see page12-18).

With this background, we see that there is one more natural vector product involving the operator  $\vec{\nabla}$ . It is the cross product of  $\vec{\nabla}$  with some vector field like  $\vec{E}$ ,  $\vec{B}$ , or  $\vec{v}$ . The cross product, for example with  $\vec{B}$ , is called the *curl* of  $\vec{B}$ .

$$\vec{\nabla} \times \vec{B} = \hat{x} (\nabla_{y} B_{z} - \nabla_{z} B_{y}) + \hat{y} (\nabla_{z} B_{x} - \nabla_{x} B_{z}) \qquad curl \qquad (8) + \hat{z} (\nabla_{x} B_{y} - \nabla_{y} B_{x})$$

With all these derivatives in the formula for  $\vec{\nabla} \times \vec{B}$ , the concept of the curl looks rather formidable. Later in this chapter we will discuss the formula for the curl in cylindrical coordinates. That formula looks even worse than Equation (8). However when we apply the curl in cylindrical coordinates to a problem with cylindrical symmetry, we end up with a simple, easily applied formula (which we will see in Equation 58).



**Figure 1** *Right hand rule for the cross product.* (*Discussed in Physics 2000, page 2-15.*)

As we have mentioned several times now, to determine a vector field we need formulas for the surface integral and the line integral. In the last chapter we saw that when we go to the small scale limit, the volume integral becomes a divergence. An example was Gauss' law which in the integral form was

$$\int \vec{E} \cdot d\vec{A} = \frac{Q_{in}}{\varepsilon_0}$$
(9)

It became the differential equation

$$\vec{\nabla} \cdot \vec{E} = \frac{\rho}{\epsilon_0} \tag{10}$$

In this chapter we will see that the differential limit of the line integral is the curl. We will see, for example, that the old form of Ampere's law (when  $\partial \vec{E}/\partial t = 0$ )

$$\oint \vec{B} \cdot d\vec{\ell} = \mu_0 i_{in} \tag{11}$$

becomes the differential equation

$$\vec{\nabla} \times \vec{B} = \mu_0 \vec{i}(x, y, z) \tag{12}$$

where  $\vec{i}(x,y,z)$  is the current density.

In our discussion of divergence, one of the important results was the divergence theorem

$$\int_{S} \vec{E} \cdot d\vec{A} = \int_{V} \vec{\nabla} \cdot \vec{E} d^{3}V \qquad \begin{array}{c} \text{divergence} \\ \text{theorem} \end{array}$$
(13)

where V is the volume bounded by a closed surface S and  $d^3V = dx dy dz$ . The divergence theorem allowed us to immediately go back and forth between surface integrals and volume integrals.

An important result of this chapter is what one could call the *curl theorem*, but which is known as *Stokes' law*. It is

$$\oint_{\substack{\text{around}\\ \text{closed path}}} \vec{B} \cdot d\vec{\ell} = \int_{\substack{\text{area of}\\ \text{closed path}}} (\vec{\nabla} \times \vec{B}) \cdot d\vec{A} \qquad Stokes' (14)$$

which relates the line integral of  $\vec{B}$  around a closed path to an integral of the curl of  $\vec{B}$  over any area bounded by the closed path. An example of a closed path is the wire loop shown in Figure (2). One of the areas bounded by this closed path is that of the soap film.

Our discussion of the curl will proceed through the remaining chapters of the text. In this chapter we will focus on deriving Stokes' theorem and applying that theorem to the theory of electricity and magnetism. This allows us to finish translating Maxwell's equations from the integral to the differential form.

In Chapter 9 we derive a set of equations called *vector identities* that simplify working with formulas involving the curl. We will use the vector identities to show that Maxwell's equations in empty space become the wave equations for electromagnetic fields.

In Chapter 11 we find that the wave equation for electromagnetic fields in the presence of electric charge and current is considerably simplified by expressing the magnetic field as the curl of a new kind of a vector field called the *vector potential*  $\vec{A}$ . This is a rather technical subject, the study of which can be put off for a while. We placed this material where we did so that you could see what happens to the electromagnetic wave equation when sources are present.

In Chapter 12 we apply the curl to the velocity field  $\vec{v}$ . It is in that chapter where you can develop the best intuitive picture of the curl. If you want to put off for a while studying the wave equation for electromagnetic fields, you can go directly from this chapter to Chapter 12 and build your intuition for curl.

In case you were wondering about Chapter 10, it deals with the extension of the continuity equation to handle compressible conserved flows, like the flow of electric charge. We discover from this work a rather remarkable result, namely that *Maxwell's equations require that electric charge be conserved*. This is one of the first completely new physical predictions we get by going to the differential form of Maxwell's equations.



**Figure 2** *Example of a surface bounded by a closed path (wire loop).* 

## STOKES' LAW

As we noted, Stokes' law, Equation (14) allows us to convert from a line integral around a closed path to a surface integral over the area bounded by the path. Once we have derived Stokes' law, it will be quite easy to use it to convert to differential equations the two Maxwell equations involving path integrals.

To derive Stokes' law, we begin by calculating the path integral of some vector field  $\vec{B}$  around a small rectangular path of sides  $\Delta x$  and  $\Delta y$  shown in Figure (3). Our arguments will be somewhat similar to those we used to derive the divergence theorem.

The line integral around the rectangle  $\Delta x \Delta y$  can be written as the four integrals

$$\oint_{\substack{\text{around}\\\Delta x \Delta y}} \vec{B} \cdot d\vec{\ell} = \int_{1}^{2} \vec{B} \cdot d\vec{\ell} + \int_{2}^{3} \vec{B} \cdot d\vec{\ell} + \int_{3}^{4} \vec{B} \cdot d\vec{\ell} + \int_{3}^{4} \vec{B} \cdot d\vec{\ell} + \int_{4}^{1} \vec{B} \cdot d\vec{\ell}$$
(15)

Along the path from point (1) to point (2), along the bottom of the rectangle, we are integrating in the x direction, thus

$$\int_{1}^{2} \vec{\mathbf{B}} \cdot d\vec{\ell} = \int_{1}^{2} \mathbf{B}_{\mathbf{x}} d\ell_{\mathbf{x}}$$
(16)

The integral of  $B_x d\ell_x$  over the bottom side can be written as

$$\int_{1}^{2} \mathbf{B}_{\mathbf{x}} d\ell_{\mathbf{x}} = \overline{\mathbf{B}}_{\mathbf{x}}(\mathbf{x}, \mathbf{y} - \Delta \mathbf{y}/2) \Delta \mathbf{x}$$
(17)

where  $\overline{B}_x(x,y-\Delta y/2)$  is the average value of  $B_x$  along the lower edge, a distance  $\Delta y/2$  below the center (x,y) of the rectangle.



Figure 3

Calculating the integral of  $\vec{B} \cdot \vec{dl}$  around a small rectangular path centered at the point (x,y).

The integral up the right hand side becomes

$$\int_{2}^{3} \vec{B} \cdot d\vec{\ell} = \int_{2}^{3} B_{y} d\ell_{y}$$
  
=  $\overline{B}_{y}(x + \Delta x/2, y) \Delta y$  (18)

where  $\overline{B}_y(x+\Delta x/2,y)$  is the average value of  $B_y$  along the right side, out at a distance  $\Delta x/2$  from the center.

On the top side, we are integrating in the –x direction, the dot product  $\vec{B} \cdot d\vec{\ell}$  is negative, and we get

$$\int_{3}^{4} \vec{B} \cdot d\vec{\ell} = \int_{3}^{4} -B_{x} d\ell_{x}$$

$$= -\overline{B}_{x}(x, y + \Delta y/2) \Delta x$$
(19)

where  $\overline{B}_x(x,y+\Delta y/2)$  is the average value of  $B_x$  on the top edge.

Going back down from point (4) to point (1) we are going in the -y direction,  $\vec{B} \cdot d\vec{\ell} = -B_y d\ell_y$  and we get

$$\int_{4}^{1} \vec{B} \cdot d\vec{\ell} = \int_{4}^{1} -B_{y} d\ell_{y}$$
  
=  $-\overline{B}_{y}(x - \Delta x/2, y) \Delta y$  (20)

Using Equations (17) through (20) in (15) gives, after some rearranging

$$\begin{split} \oint \vec{B} \cdot d\vec{\ell} &= \left[ \frac{\overline{B}_{y}(x + \Delta x/2, y) - \overline{B}_{y}(x - \Delta x/2, y)}{\Delta x} \right] \Delta x \Delta y \\ &- \left[ \frac{\overline{B}_{x}(x, y + \Delta y/2) - \overline{B}_{x}(x, y - \Delta y/2)}{\Delta y} \right] \Delta x \Delta y \end{split}$$

$$(21)$$

As a first approximation to Equation (21), we could replace the average values of  $\overline{B}_x$ ,  $\overline{B}_y$  on the four sides by the actual values of  $B_x$ ,  $B_y$  at the center of each side. For example, since the center of the side from (2) to (3) is at the point (x+ $\Delta x/2$ ,y), we would be making the substitution for that side of

$$\overline{B}_{v}(x + \Delta x/2, y) \to B_{v}(x + \Delta x/2, y)$$
(22)

I.e., we would be removing the bars over the values of B in Equation (21).

When we remove the bars and then take the limit as  $\Delta x \rightarrow 0$  and  $\Delta y \rightarrow 0$ , the first square bracket in Equation (21) becomes the partial derivative of  $B_y$  with respect to x

$$\lim_{\Delta x \to 0} \left[ \frac{B_{y}(x + \Delta x/2, y) - B_{y}(x - \Delta x/2, y)}{\Delta x} \right] = \frac{\partial B_{y}}{\partial x}$$
(23)

and the second square bracket in Equation (21) becomes  $\partial B_x/\partial y$ . In this approximation, Equation (21) becomes

$$\oint_{\substack{\text{around}\\\Delta x \Delta y}} \vec{B} \cdot d\vec{\ell} = \left[ \frac{\partial B_y}{\partial x} - \frac{\partial B_x}{\partial y} \right] (\Delta x \Delta y) \quad (24)$$

The approximation we made to get Equation (24), which was replacing the average value of B along a line by the value at the center of the line, assumes that variations along the line (e.g. changes in  $B_x$  in the x direction) are not as important as variations perpendicular to the line (e.g. changes in  $B_x$  in the y direction). This is somewhat similar to the situation we had in our derivation of the divergence theorem where changes in the field were important in one direction and not in the other.

In the appendix to Chapter 7 we used a Taylor series expansion to show that as  $\Delta x$ ,  $\Delta y$  or  $\Delta z$  went to zero, the variations we ignored went to zero faster than the variations we kept. They were proportional to a higher power of  $\Delta x$ ,  $\Delta y$  or  $\Delta z$ , and therefore did not contribute in the calculus limit.

We leave it as an exercise for the ambitious reader to show, using arguments similar to those made in the appendix to Chapter 7, that by replacing average values  $\overline{B}_x$  and  $\overline{B}_y$  by center values  $B_x$  and  $B_y$ , we are making errors that go to zero faster than the terms we keep. I.e., show that the errors are of the order  $\Delta x$ ,  $\Delta y$  or  $\Delta z$  smaller than the terms we keep.

With Equation (24), we have the formula for the line integral around one small rectangle lying in the xy plane. We can generalize this result by turning the area element  $(\Delta x \Delta y)$  into a vector  $\overrightarrow{\Delta A}$ . An area vector  $\overrightarrow{\Delta A}$  is perpendicular to the surface as shown in Figure (4). In this case, where the surface is in the xy plane, we see that  $\overrightarrow{\Delta A}$  is purely z directed, and we can write  $\Delta x \Delta y = (\Delta A)_z$ . With this notation Equation (24) becomes

$$\oint_{\substack{\text{around}\\\Delta A}} \vec{B} \cdot d\vec{\ell} = \left[ \frac{\partial B_y}{\partial x} - \frac{\partial B_x}{\partial y} \right] (\Delta A)_z$$
(25)

Next, we notice that the z component of the curl of  $\vec{B}$  is given by Equation (8) as

$$(\vec{\nabla} \times \vec{B})_z = (\nabla_x B_y - \nabla_y B_x) = \frac{\partial B_y}{\partial x} - \frac{\partial B_x}{\partial y}$$
 (26)

so that Equation (25) becomes

$$\oint_{\substack{\text{around}\\ \Delta A}} \vec{B} \cdot d\vec{\ell} = (\vec{\nabla} \times \vec{B})_z (\Delta A)_z$$
(27)

The obvious extension of Equation (27) to the case where our area  $\Delta A$  does not happen to lie in the xy plane, where the vector  $\overline{\Delta A}$  has components other than  $(\Delta A)_z$ , is to recognize that in Equation (27) we are looking at one term in the vector dot product

$$\oint_{\substack{\text{around}\\ \Delta A}} \vec{B} \cdot d\vec{\ell} = (\vec{\nabla} \times \vec{B}) \cdot \vec{\Delta} \vec{A}$$
(28)

### Exercise 1







With Equation (28), we have the formula for the line integral around a small rectangular area  $\overrightarrow{\Delta A}$  of any orientation. The final step is to determine the line integral around a finite loop like the wire loop with the soap film across it, shown in Figure (1).

The way we can do this is to conceptually cut the soap film up into many tiny rectangles as shown in Figure (6). Think of the soap film as being replaced by a window screen, with the rectangles being the holes in the window screen.

At each hole, each rectangle, we have a vector  $\overrightarrow{\Delta A_i}$  that is oriented perpendicular to the surface as shown in Figure (7). The positive direction is determined by noting which way we are going around the loop, and then using the right hand rule. For Figure (6), the positive direction is up out of the paper.

Next we note that when two rectangles touch each other, the part of the line integrals on the touching sides cancel, and we are left with a line integral around the perimeter of the two rectangles as shown in Figure (8).



Applying this argument to all rectangles in Figure (6), we see that when we add up the line integrals for all the rectangles, we end up with the line integral around the outside perimeter of the surface. Mathematically we can write this as

$$\oint_{\substack{\text{around}\\\text{whole}\\\text{surface}}} \vec{B} \cdot d\vec{\ell} = \frac{\text{sum of the line}}{\substack{\text{integrals around}\\\text{each small}}} \qquad (29)$$

Using Equation (28) for the line integral around  $\overrightarrow{\Delta A}_i$  we get

$$\oint \vec{B} \cdot d\vec{\ell} = \sum_{i} (\vec{\nabla} \times \vec{B}) \cdot \vec{\Delta A}_{i}$$
(30)
around
whole
surface

Taking the limit as the  $\overrightarrow{\Delta A}_i$  goes to zero turns this sum into an integral, giving

$$\oint_{\substack{\text{around}\\\text{perimeter of}\\\text{a surface S}}} \vec{B} \cdot d\vec{\ell} = \int_{\substack{\text{over the}\\\text{surface S}}} (\vec{\nabla} \times \vec{B}) \cdot d\vec{A}$$
Stokes' law
(31)

which is Stokes' law. It says that we get the line integral of any vector field  $\vec{B}$  around the perimeter of a surface S by integrating the flux of  $(\vec{\nabla} \times \vec{B})$  out through the surface.



## When two rectangles touch, the line integrals on the paths between them cancel, leaving a line integral around the perimeter of the two rectangles.

In the future we will shorten our notation by letting C be some closed path, and the surface S be a surface like our soap film, that is bounded by the path. Then we simply write

$$\oint_{C} \vec{B} \cdot d\vec{\ell} = \int_{S} (\vec{\nabla} \times \vec{B}) \cdot d\vec{A}$$
 Stokes' law

(31a)

Our use of the soap film analogy for the surface S is important for it emphasizes the fact that there is no one correct surface. Just as you can change the shape of a soap film by gently blowing on it (don't blow a bubble), you can use different surfaces S as long as they are bounded by the same circuit C.

We also want to emphasize that the quantity  $(\vec{\nabla} \times \vec{B})$ is itself a vector field, and that the integral of  $(\vec{\nabla} \times \vec{B}) \cdot d\vec{A}$  over a surface is the flux of  $(\vec{\nabla} \times \vec{B})$ through that surface. Thus, we should remember Stokes' law as telling us that *the line integral of*  $\vec{B}$ *around the circuit C is equal to the flux of*  $(\vec{\nabla} \times \vec{B})$ *through the circuit C*.



**Figure 9** When the current flows at an angle  $\theta$  as shown, the total current through  $\Delta A$  is  $i(x,y,z) \Delta A \cos \theta$ .

## AMPERE'S LAW

The original form of Ampere's law, before Maxwell's addition of the  $\partial \Phi_E / \partial t$  term, was given in Chapter 29 of the Physics text as

$$\oint \vec{B} \cdot d\vec{\ell} = \mu_0 I_{\text{enclosed}}$$
(29-26)  
any closed  
path

It says that the line integral of  $\vec{B}$  around any closed path is equal to  $\mu_0$  times the total current flowing through that path. Since Stokes' law tells us that the line integral of  $\vec{B}$  around any closed path is equal to the total flux of ( $\vec{\nabla} \times \vec{B}$ ) through that path, there must be a close relationship between the vector field ( $\vec{\nabla} \times \vec{B}$ ) and the electric current. That is the relationship we want to establish.

The first step is to express the total current i through a closed path in terms of the current density  $\vec{i}(x,y,z)$ . The current density  $\vec{i}(x,y,z)$  is a vector field whose direction at each point in space is the direction of flow on the electric current  $\vec{i}$  there, and whose magnitude is equal to the density of current, which has the dimensions of the number of amperes per square meter.

Calculating the electric current through a small area element  $\overline{\Delta A}$  is analogous to calculating the flux of water through an area element  $\overline{\Delta A}$ , a calculation we did in Equation (3) of Chapter 29 of the Physics text. From Figure (9), you can see that the current through  $\Delta A$  will be a maximum, will have the value  $\vec{i}(x,y,z)\Delta A$  when the area  $\Delta A$  is perpendicular to the flow. This is when the vector  $\overrightarrow{\Delta A}$  is parallel to  $\vec{i}(x,y,z)$ . For any other orientation of  $\overrightarrow{\Delta A}$ , the current  $\Delta I$  through  $\overrightarrow{\Delta A}$  will be equal to  $\vec{i}(x,y,z)\Delta A \cos\theta$  which is equal to the dot product of the vectors  $\vec{i}(x,y,z)$  and  $\overrightarrow{\Delta A}$ . Thus

$$\Delta I = \vec{i}(x,y,z) \cdot \overrightarrow{\Delta A} = \frac{\text{current through an}}{\text{area element } \overrightarrow{\Delta A}}$$
(32)

To calculate the total current  $I_{enclosed}$  through an entire surface S, we break the surface up into small areas  $\overline{\Delta A}_i$  as we did in Figure (6), calculate the current  $\Delta I_i$  through each  $\overline{\Delta A}_i$ , and add up all the  $\Delta I_i$  to get the total.

$$I_{\text{enclosed}} = \sum_{i} \Delta I_{i} = \sum_{i} \vec{i} (x_{i}, y_{i}, z_{i}) \cdot \Delta \vec{A}_{i} \quad (33)$$

Taking the limit as the  $\overrightarrow{\Delta A}_i$  go to zero size gives us the surface integral

$$I_{enclosed} = \int \vec{i}(x,y,z) \cdot d\vec{A} \qquad \begin{array}{c} total \ current \\ through \ a \\ closed \ path \ C \end{array} (34)$$

Using our new formula for I<sub>enclosed</sub> in Ampere's law, Equation (29-26), gives

$$\oint_{any} \vec{B} \cdot d\vec{\ell} = \mu_0 \int_{over the area bounded by the closed path} \vec{i}(x,y,z) \cdot d\vec{A}$$
(35)

Following a procedure similar to the one we used in our discussion of Gauss' law in Chapter 7, we will use Stokes' law to convert the line integral of  $\vec{B}$  to a surface integral, so that both terms in Ampere's law are surface integrals. With

$$\oint_{\mathbf{C}} \vec{\mathbf{B}} \cdot d\vec{\ell} = \int_{\mathbf{S}} (\vec{\nabla} \times \vec{\mathbf{B}}) \cdot d\vec{\mathbf{A}}$$
(31) repeated

Equation (35) becomes

$$\int_{\substack{\text{surface}\\S}} (\vec{\nabla} \times \vec{B}) \cdot d\vec{A} = \int_{\substack{\text{surface}\\S}} \mu_0 \vec{i}(x,y,z) \cdot d\vec{A} \quad (36)$$

where we took the constant  $\mu_0$  inside the integral. The surfaces for the two integrals only have to have the same perimeter C, but we are free to choose identical surfaces, and thus combine the two integrals into one giving

$$\int_{\substack{\text{any} \\ \text{surface} \\ S}} \left[ (\vec{\nabla} \times \vec{B}) - \mu_0 \vec{i} (x, y, z) \right] \cdot d\vec{A} = 0 \quad (37)$$

We then argue that if Equation (37) is to hold for any surface S, the only way for that to happen is to set the integrand, the stuff in the square brackets, equal to zero, giving

$$\vec{\nabla} \times \vec{B} = \mu_0 \vec{i}(x, y, z)$$
(38)

Equation (38) is the differential form of the original Ampere's law

$$\oint \vec{B} \cdot d\vec{\ell} = \mu_0 I_{\text{enclosed}} \qquad (29-26) \text{ repeated}$$

In Chapter 32 of the Physics text we explained why Maxwell added a term to Ampere's law to get

$$\oint_{\substack{\text{around } a \\ \text{closed} \\ \text{circuit } C}} \vec{B} \cdot d\vec{\ell} = \mu_0 I_{enclosed} + \mu_0 \epsilon_0 \frac{d\Phi_E}{dt}$$
(32-11)

where  $\Phi_{\rm E}$ , the electric flux through the closed circuit is given by

$$\Phi_{\rm E} = \int_{\rm S} \vec{\rm E} \cdot d\vec{\rm A} \tag{39}$$

and S is any surface bounded by the closed circuit C.

To include the  $d\Phi_E/dt$  term in our differential form of Ampere's law, we need to evaluate

$$\frac{d}{dt}\Phi_{\rm E}(t) = \frac{d}{dt}\int_{\rm S} \vec{\rm E}(x,y,z,t) \cdot d\vec{\rm A} \tag{40}$$

where the field E is not only a function of space (x,y,z) but also of time (t).

On the left side of Equation (40) we have  $d\Phi_E(t)/dt$ which is simply the time derivative of some function  $\Phi_E(t)$  of time. That is a straightforward derivative. On the right, we have the derivative of the integral of a quantity  $\vec{E}(x,y,z,t)$  which is a function of four variables. What we are going to do this one time, is to be very careful about how we bring the time derivative inside the integral, and see what we get when we do. Our first step will be to write the integral over the surface as the sum over many small but finite areas  $\Delta A_i$ 

$$\frac{d\Phi_E}{dt} = \frac{d}{dt} \sum_{i} \vec{E}(x_i, y_i, z_i, t) \cdot \Delta \vec{A}_i$$
(41)

where  $(x_i, y_i, z_i)$  is the coordinate of the area element  $\Delta A_i$ . By working with a sum of finite terms, we can see that the change in time of the sum will be the sum of the changes in each term

$$\frac{d\Phi_{E}}{dt} = \sum_{i} \frac{d}{dt} \left[ \vec{E}(x_{i}, y_{i}, z_{i}, t) \cdot \vec{\Delta A}_{i} \right]$$
(42)

During this calculation, we are keeping the surface S and all the  $\Delta \vec{A}_i$  fixed. At any given  $\Delta \vec{A}_i$  the only thing that is allowed to change is the field  $\vec{E}$  at the point  $(x_i, y_i, z_i)$ . Thus we have

$$\frac{d\Phi_{\rm E}}{dt} = \sum_{\rm i} \left[ \frac{d\vec{\rm E}(x_{\rm i}, y_{\rm i}, z_{\rm i}, t)}{dt} \right] \cdot \Delta \vec{\rm A}_{\rm i}$$
(43)

The term in the square brackets is the change in the variable  $\vec{E}(x,y,z,t)$  as we change the time (t) while holding the other three variables constant at  $x = x_i$ ,  $y = y_i$ ,  $z = z_i$ . This is precisely what we mean by the *partial derivative* of  $\vec{E}(x,y,z,t)$  with respect to (t).

$$\frac{d}{dt}\vec{E}(x_{i},y_{i},z_{i},t) = \frac{\partial\vec{E}(x,y,z,t)}{\partial t}\bigg|_{\substack{x=x_{i}\\y=y_{i}\\z=z_{i}}}$$
(44)

Thus we have

$$\frac{d\Phi_{E}}{dt} = \sum_{i} \frac{\partial \vec{E}(x,y,z,t)}{\partial t} \bigg|_{\substack{x = x_{i} \\ y = y_{i} \\ z = z_{i}}} \Delta \vec{A}_{i}$$
(45)

We can now go back to the limit as  $\Delta \vec{A}_i$  goes to zero, giving

$$\frac{d\Phi_{\rm E}}{dt} = \int_{\rm S} \frac{\partial \vec{\rm E}(x,y,z,t)}{\partial t} \cdot d\vec{\rm A}$$
(46)

Writing  $d\Phi_{\rm F}/dt$  in Equation (46) as an integral gives

$$\frac{d}{dt} \int_{\substack{\text{fixed}\\\text{surface}\\S}} \vec{E}(x,y,z,t) \cdot d\vec{A} = \int_{\substack{\text{fixed}\\\text{surface}\\S}} \frac{\partial \vec{E}(x,y,z,t) \cdot d\vec{A}}{\partial t}$$
(47)

In writing Equation (47) we placed special emphasis on the fact that the surface S (and also the  $\Delta A_i$ 's) were fixed, did not change with time. Later, in the first fluid dynamics chapter, we will want to calculate the rate of change of flux through a moving surface. (In that case it will be a surface that moves with the fluid particles.) When we allow the surface S to move, then in going from Equation (42) to (43), we get more terms representing changes in the  $\Delta A_i$ .

But with the fixed surface, Equation (47) tells us that we can bring the time derivative inside the integral if we change the derivative to a partial derivative with respect to time.

## **Exercise 2**

Start from the integral form of Ampere's law

$$\oint \vec{B} \cdot d\vec{\ell} = \mu_0 |_{\text{enclosed}} + \mu_0 \epsilon_0 \frac{d\Phi_E}{dt}$$
(32-11)

Using Equation (39) for  $\Phi_{\rm E}$ , and using Equation (47), show that the corresponding differential equation is

$$\nabla \times \vec{\mathsf{B}} = \mu_0 \vec{\mathsf{i}} + \mu_0 \varepsilon_0 \frac{\partial \vec{\mathsf{E}}}{\partial t}$$
(48)

### Exercise 3

As a review, start with all of Maxwell's equations in integral form, as summarized in Equation (32-19) of the Physics text

$$\int \vec{E} \cdot d\vec{A} = \frac{Q_{in}}{\varepsilon_0} \qquad \text{Gauss' law}$$
closed surface
$$\int \vec{B} \cdot d\vec{A} = 0 \qquad \text{no monopole}$$

closed surface

$$\begin{split} \oint \vec{B} \cdot d\vec{\ell} &= \mu_0 I + \mu_0 \epsilon_0 \frac{d\Phi_E}{dt} \text{ Ampere's law} \quad (32\text{-}19) \\ \oint \vec{E} \cdot d\vec{\ell} &= \frac{-d\Phi_B}{dt} \quad \text{Faraday's law} \end{split}$$

Faraday's law

and show that in differential form, the equations are

$$\vec{\nabla} \cdot \vec{E} = \frac{\rho}{\epsilon_0} \qquad \text{Gauss' law}$$

$$\vec{\nabla} \cdot \vec{B} = 0 \qquad \text{no monopole}$$

$$\vec{\nabla} \times \vec{B} = \mu_0 \vec{i} + \mu_0 \epsilon_0 \frac{\partial \vec{E}}{\partial t} \qquad \text{Ampere's law} \qquad (49)$$

$$\vec{\nabla} \times \vec{E} = -\frac{\partial \vec{B}}{\partial t} \qquad \text{Faraday's law}$$

## **CURL OF THE MAGNETIC FIELD OF A WIRE**

In the section after this, we will discuss the formula for the curl in cylindrical coordinates, a rather formidable looking formula. We will then apply it to the calculation of the curl  $\vec{\nabla} \times \vec{B}$  of the magnetic field of a straight wire. A lot of terms are involved but, most of them go to zero and we are left with what appears to be a surprisingly simple result. The result should be no surprise however, if we first look at Ampere's law in differential form, as applied to the field of a wire.

The magnetic field produced by a steady current in a wire was shown in Figure (28-14) in Chapter 28 of the Physics text. The current (i) is confined to the wire, and the magnetic field travels in circles around the wire. If the current density is more or less uniform in the wire, then we have a circular magnetic field inside the wire also (a field you calculated in Exercise 4 of Chapter 29). The result is sketched in Figure (10). For a steady current, where  $\partial \vec{E}/\partial t = 0$ , Ampere's law in differential form is simply

$$\vec{\nabla} \times \vec{B} = \mu_0 \vec{i}(x,y,z)$$
 (38) repeated

The first thing to note about Equation (38) is that in all places where the current density  $\vec{i}(x,y,z)$  is zero, the curl  $\vec{\nabla} \times \vec{B}$  must also be zero. Since the current is confined to the wire,  $\vec{\nabla} \times \vec{B}$  must be confined there, and *the curl of the magnetic field outside the wire must be zero*. It will take us several pages to obtain the same result using the formulas for the curl.

Next we note that the current density  $\vec{i}(x,y,z)$  is not only confined to the wire, but also directed along the wire. Thus  $\vec{\nabla} \times \vec{B}$  must not only be confined to the wire, **but also directed along the wire** as shown in Figure (11). As a result we know what  $\vec{\nabla} \times \vec{B}$  must look like before we do any calculations.

In the next sections we will go through the calculation of the curl of this magnetic field. When we finally get the simple results described above, you can look upon that as a check that the formulas for curl are correct after all.



**Figure 10** *The magnetic field inside and outside a wire carrying a uniform current.* 



**Figure 11** The curl of that magnetic field, determined by  $\vec{\nabla} \times \vec{B} = \mu_0 \vec{i}(x,y,z)$ .

## **CURL IN CYLINDRICAL COORDINATES**

In our study of the gradient in Chapter 3 and of Schrödinger's equation in Chapter 6, we saw that when a problem had cylindrical or spherical symmetry, there was a considerable advantage to using the formulas in cylindrical or spherical coordinates. Very often problems involving the curl, like the magnetic field of the current in a straight wire, have a cylindrical symmetry. For such problems it is much easier to work with the curl in cylindrical coordinates.

Deriving formulas for curl  $\vec{\nabla} \times \vec{B}$  and divergence  $\vec{\nabla} \cdot \vec{E}$  in cylindrical or spherical coordinates is made difficult because of the unit vectors. In Cartesian coordinates, the unit vectors are constant. But in other coordinate systems the unit vectors change as we move around in space. When we take the partial derivative of a vector, we also have to include the effects of changes in the unit vectors.

In the appendix to Chapter 4, where we calculated  $\vec{\nabla} \cdot (\vec{\nabla} f) = \nabla^2 f$  in spherical polar coordinates, most of the calculation dealt with the changing unit vectors. In a more closely related example, suppose we have the vector  $\vec{B}$  expressed in cylindrical coordinates as

$$\vec{B} = \hat{r}B_r + \hat{\theta}B_{\theta} + \hat{z}B_z$$

where the unit vectors  $\hat{r}$ ,  $\hat{\theta}$ , and  $\hat{z}$  are shown in Figure (12). If we make a change in the angle  $\theta$  from  $\theta$  to  $\theta + \Delta \theta$ , the unit vectors  $\hat{r}$  and  $\hat{\theta}$  change directions by an angle  $\Delta \theta$  as shown in Figure (13).



**Figure 12** *The unit vectors in cylindrical coordinates.* 

When we calculate the partial derivative of the vector  $\vec{B}$ , as we change the angle  $\theta$  from  $\theta$  to  $\theta + \Delta \theta$ , we not only have to include the change in the value of  $\vec{B}$  as we move from points (1) to (2) in Figure (11), we also have to account for the fact that the unit vectors  $\hat{r}$  and  $\hat{\theta}$  have also changed. This change mixes up the components of  $\vec{B}$ .

It is not impossible to work out the formulas for the divergence or curl of a vector in cylindrical or spherical coordinates, but one is not likely to do it on the back of an envelope and get the right answer. Any practicing physicist or engineer, who needs to use these formulas, looks them up in a reliable reference.

What we will do is simply state the formula for curl in cylindrical coordinates, and then check that the formula gives the simple results we discussed in the last section for the case of the magnetic field of a wire. At the end of this text, in the *Formulary*, we summarize all the formulas for gradient, divergence and curl, in Cartesian, cylindrical and spherical coordinates. Such a summary can be a very useful thing to have.

Given a field  $\vec{B}$  expressed in cylindrical coordinates as

$$\vec{\mathbf{B}} = \hat{\mathbf{r}}\mathbf{B}_{r} + \hat{\mathbf{\theta}}\mathbf{B}_{\theta} + \hat{\mathbf{z}}\mathbf{B}_{z}$$
(50)

the formula for the curl is

$$(\vec{\nabla} \times \vec{B})_{r} = \frac{1}{r} \frac{\partial B_{z}}{\partial \theta} - \frac{\partial B_{\theta}}{\partial z}$$
$$(\vec{\nabla} \times \vec{B})_{\theta} = \frac{\partial B_{r}}{\partial z} - \frac{\partial B_{z}}{\partial r}$$
$$(\vec{\nabla} \times \vec{B})_{z} = \frac{1}{r} \frac{\partial}{\partial r} (rB_{\theta}) - \frac{1}{r} \frac{\partial B_{r}}{\partial \theta}$$
(51)





## CALCULATING THE CURL OF THE MAGNETIC FIELD OF A WIRE

While Equation (51) for  $\vec{\nabla} \times \vec{B}$  in cylindrical coordinates looks worse than the curl in Cartesian coordinates, you will see a major simplification when applied to a problem with cylindrical symmetry. The magnetic field of a wire travels in circles about the wire as shown in Figure (14). We see that  $\vec{B}$  has only a  $\hat{\theta}$  component  $B_{\theta}$ . In addition, the value of  $B_{\theta}$  does not depend on, i.e., change with, the height z or the angle  $\theta$ . Thus we can write  $\vec{B}$  as

$$\vec{\mathbf{B}} = \hat{\theta} \mathbf{B}_{\theta}(\mathbf{r}) \qquad \begin{array}{l} magnetic \ field \ of \\ a \ straight \ wire \end{array} \tag{52}$$

where the only variable  $B_{\theta}$  depends upon is the radius.

## **Outside the wire**

We will first calculate  $\vec{B}$  using the integral form of Gauss' law, and then see what happens when we apply the curl formula, Equation (51) to  $\vec{B}$ . Integrating  $\vec{B}$  around the circular path of radius r, shown by the dotted circle in Figure (12) gives



This is the result we saw in Chapter 28 of the Physics text. Here  $i_{enclosed}$  is equal to the total current  $i_{tot}$  because our path goes around the wire.

We are now ready to plug in the values

$$B_{\rm r} = 0$$
  

$$B_{\theta} = \frac{\mu_0 i_{\rm tot}}{2\pi r}$$
  

$$B_{\rm z} = 0$$
(54)

into Equation (51) to get the value of the curl

$$\vec{\nabla} \times \vec{B} = \hat{r} (\vec{\nabla} \times \vec{B})_{r} + \hat{\theta} (\vec{\nabla} \times \vec{B})_{\theta} + \hat{z} (\vec{\nabla} \times \vec{B})_{z} \quad (55)$$

Because  $B_r$  and  $B_z$  are zero, a lot of the terms in the formula for  $\vec{\nabla} \times \vec{B}$  vanish, and we are left with

$$(\vec{\nabla} \times \vec{B})_{r} = -\frac{\partial B_{\theta}}{\partial z}$$
$$(\vec{\nabla} \times \vec{B})_{\theta} = 0$$
$$(\vec{\nabla} \times \vec{B})_{z} = \frac{1}{r} \frac{\partial}{\partial r} (r B_{\theta})$$
(56)

You should check for yourself that this is all that is left of  $\vec{\nabla} \times \vec{B}$  for the  $\vec{B}$  of Equation (54).

We now note that  $B_{\theta}(r) = \mu_0 i_{tot} / 2\pi r$  depends only on the variable r and has no z dependence. Thus

$$\frac{\partial B_{\theta}(\mathbf{r})}{\partial z} = 0 \tag{57}$$

and all we are left with for the curl is

$$(\vec{\nabla} \times \vec{B})_{z} = \frac{1}{r} \frac{\partial}{\partial r} (rB_{\theta})$$
(58)

Equation (58) applies to any vector field that looks like the magnetic field in Figure (12). It applies to any vector field of the form

$$\vec{\mathbf{B}} = \hat{\boldsymbol{\theta}} \mathbf{f}(\mathbf{r}) \tag{59}$$

where f(r) is any function of r. These are the kinds of fields we are most likely to deal with in a discussion of the curl, in which case we can use the much simpler Equation (58). Applying Equation (58) to our special value  $B_{\phi} = \mu_0 i_{tot} / 2\pi r$ , we get

$$\left( \vec{\nabla} \times \vec{\mathbf{B}} \right)_{z} = \frac{1}{r} \frac{\partial}{\partial r} \left( r \mathbf{B}_{\theta} \right)$$

$$= \frac{1}{r} \frac{\partial}{\partial r} \left[ r \left( \frac{\mu_{0} \mathbf{i}_{tot}}{2\pi r} \right) \right]$$
(60)

Notice that the r's in the square bracket cancel, leaving us with

$$\left(\vec{\nabla} \times \vec{B}\right)_{z} = \frac{1}{r} \frac{\partial}{\partial r} \left(\frac{\mu_{0} i_{tot}}{2\pi}\right)$$
(61)

We see that  $\mu_0 i_{tot}/2\pi$  is a constant and the derivative of a constant is zero

$$\frac{\partial}{\partial \mathbf{r}} \left( \frac{\mu_0 \mathbf{i}_{\text{tot}}}{2\pi} \right) = 0 \tag{62}$$

Thus we end up with the simple result

$$\vec{\nabla} \times \vec{B} = 0$$
  $\left( \text{for } B_{\theta} = \frac{\mu_0 i_{\text{tot}}}{2\pi r} \right)$  (63)

This is what we expected from our earlier discussion of Ampere's law in differential form. Neglecting the  $\partial \vec{E}/\partial t$  term, the law is

$$\vec{\nabla} \times \vec{B} = \mu_0 \vec{i}(x,y,z)$$
 (38) repeated

where the vector  $\vec{i}(x,y,z)$  is the current density. Since the current is confined to the wire, the curl  $\vec{\nabla} \times \vec{B}$  must also be confined to the wire, and be zero outside.

## **Inside the Wire**

What about inside the wire where the current density is not zero? Equation (53) does not apply there because the formula  $B_{\theta} = \mu_0 i_{tot}/2\pi r$  applies only outside the wire.

To calculate the magnetic field inside the wire, we have to know something about the current density. Let us assume that we have a uniform current inside a wire of radius R. We will apply Ampere's law to a circular path of radius r as shown in the end view of the wire in Figure (15).

The amount of current enclosed by our path of radius r is, for a uniform current, simply the total current  $i_{tot}$  times the ratio of the area  $\pi r^2$  of the path, to the area  $\pi R^2$  of the wire

$$i_{\text{enclosed}} = i_{\text{total}} \frac{\pi r^2}{\pi R^2} = i_{\text{tot}} \frac{r^2}{R^2}$$
(64)

Using this value in Ampere's law, we get for the magnetic field inside the wire

$$\oint \vec{B} \cdot d\vec{\ell} = \mu_0 i_{\text{enclosed}}$$

$$\left(B_\theta \times 2\pi r\right) = \mu_0 i_{\text{tot}} \frac{r^2}{R^2}$$
(65)

One of the r's cancels, and we are left with

$$\mathbf{B}_{\theta}(\mathbf{r}) = \left[\frac{\mu_0 \mathbf{i}_{\text{total}}}{2\pi R^2}\right] \mathbf{r}$$
(66)

where everything in the square brackets is a constant. You derived this result in Exercise (29-4) of the Physics text.



Calculating the magnetic field inside the wire, assuming a uniform current density.

Repeating Equation (66), we had for the field inside the wire

$$B_{\theta}(\mathbf{r}) = \left[\frac{\mu_0 i_{\text{total}}}{2\pi R^2}\right] \mathbf{r}$$
(66)

We see that  $B_{\theta}$  increases linearly with r until we reach the surface of the wire at r=R, as shown in Figure (16). Then outside the wire,  $B_{\theta}$  drops off as 1/r.

To simplify the formulas, let us write  $B_{\boldsymbol{\theta}}$  inside the wire as

$$B_{\theta}(r) = kr \qquad \frac{inside}{wire} \qquad (66a)$$

where

$$k = \frac{\mu_0 i_{\text{total}}}{2\pi R^2} \tag{66b}$$

The curl of this value of  $B_{\theta}$  is given by Equation (58) as

$$(\vec{\nabla} \times \vec{B})_{z} = \frac{1}{r} \frac{\partial}{\partial r} (rB_{\theta})$$
$$= \frac{1}{r} \frac{\partial}{\partial r} (rkr)$$
$$= \frac{k}{r} \frac{\partial}{\partial r} (r^{2})$$
(67)

Since  $\partial(\mathbf{r}^2)/\partial \mathbf{r} = 2\mathbf{r}$ , we get

$$(\vec{\nabla} \times \vec{B})_z = \frac{k}{r}(2r) = 2k \tag{68}$$

Putting back our value for  $k = \mu_0 i_{tot}/2\pi R^2$  we get

$$(\vec{\nabla} \times \vec{B})_z = \mu_0 \left( \frac{i_{tot}}{\pi R^2} \right)$$
(69)

Now  $i_{tot}/\pi R^2$  is the total current in the wire divided by the area of the wire, which is the current density i(x,y,z). Since the current is  $\hat{z}$  directed, we can write the current density as

$$\vec{i}(x,y,z) = \hat{z} \frac{i_{tot}}{\pi R^2}$$
(70)

and Equation (70) can be written as the vector equation

$$\vec{\nabla} \times \vec{B} = \mu_0 \vec{i}(x,y,z)$$
 (38) repeated

which is the differential form of Ampere's law (for  $\partial \vec{E} / \partial t = 0$ ).

This is the result we expected in the first place. The fact that we got back to Ampere's law serves as a check that the formulas for the curl in cylindrical coordinates are working.



The magnetic field inside and outside the wire, for a uniform current density inside the wire.

## Calculus 2000-Chapter 9 Electromagnetic Waves

In the Physics text we had some difficulty showing that Maxwell's equations led to the prediction of the existence of electromagnetic radiation. The problem was that the integral form of Maxwell's equations are not particularly well suited for the derivation. The best we could do was to show that the wave pulse, shown in Figure (32-16) reproduced here, travels out at a speed  $v = 1/\sqrt{\mu_0 \varepsilon_0}$  which turns out to be the speed of light. In discussing light waves, we made the argument that if we started with a series of wave pulses shown in Figure (32-23a) and smoothed them out, we could get the sinusoidal pulse shown in (32-23b). We never did show that the smoothed out version was actually a solution of Maxwell's equations, or that the sinusoidal structure traveled at a speed  $c = 1/\sqrt{\mu_0 \varepsilon_0}$ . With the differential form of Maxwell's equations, we can now do that.



**Figure 32-16** Electromagnetic pulse produced by turning the current on and then quickly off. We will see that this structure agrees with Maxwell's equations.



Figure 32-23 by smoot Structure of electric and magnetic fields in light and radio waves.

## **VECTOR IDENTITIES**

To use the differential forms of Maxwell's equations, it is convenient to first develop three formulas known as *vector identities*. These are mathematical relationships involving curls that apply to any vector field. We will state these identities first and then spend the rest of the section deriving them. You should go through these derivations at least once to get a feeling for how they work and how general they are.

## **Identity 1**

The curl of a gradient  $\vec{\nabla} f$  is zero for any scalar field f(x,y,z).

$$\vec{\nabla} \times (\vec{\nabla} f) = 0 \tag{1}$$

## **Identity 2**

The divergence of a curl is zero. That is, for any vector field  $\vec{A}(x,y,z)$ 

$$\vec{\nabla} \cdot (\vec{\nabla} \times \vec{A}) = 0 \tag{2}$$

#### **Identity 3**

This identity gives us a formula for the curl of a curl. The formula is

$$\vec{\nabla} \times (\vec{\nabla} \times \vec{A}) = -(\vec{\nabla} \cdot \vec{\nabla})\vec{A} + \vec{\nabla}(\vec{\nabla} \cdot \vec{A})$$
(3)

where  $\vec{\nabla} \cdot \vec{\nabla} = \nabla_x \nabla_x + \nabla_y \nabla_y + \nabla_z \nabla_z$  is the Laplacian operator discussed in Chapter 4. We will often use the notation

$$\vec{\nabla} \cdot \vec{\nabla} \equiv \nabla^2 = \nabla_x \nabla_x + \nabla_y \nabla_y + \nabla_z \nabla_z \qquad (4)$$

so that the vector identity can be written as

$$\vec{\nabla} \times (\vec{\nabla} \times \vec{A}) = -\nabla^2 \vec{A} + \vec{\nabla} (\vec{\nabla} \cdot \vec{A})$$
(5)

In the special case that  $\vec{A}$  has zero divergence, if  $\vec{\nabla}\cdot\vec{A}=0$  , then we get

$$\vec{\nabla} \times (\vec{\nabla} \times \vec{A}) = -\nabla^2 \vec{A} \qquad if \ \vec{\nabla} \cdot \vec{A} \\ is \ zero$$
(5a)

## **Proof of Identity 1**

The proof of these identities relies on the fact that we can interchange the order of partial differentiation, a result we prove in the appendix to this chapter. As an example of how this is used, consider one component of the first identity. Using the cross product formula

$$(\vec{A} \times \vec{B})_{x} = A_{y}B_{z} - A_{z}B_{y}$$
(6)

we get

$$\begin{bmatrix} \vec{\nabla} \times (\vec{\nabla} f) \end{bmatrix}_{x} = \nabla_{y} (\nabla_{z} f) - \nabla_{z} (\nabla_{y} f) \\ = \nabla_{y} \nabla_{z} f - \nabla_{z} \nabla_{y} f$$
(7)

Interchanging  $\nabla_y \nabla_z$  to get  $\nabla_y \nabla_z f = \nabla_z \nabla_y f$  immediately makes this component zero. The same thing happens to the y and z components of  $\vec{\nabla} \times (\vec{\nabla} f)$ , thus the entire expression is zero.

## **Proof of Identity 2**

To prove the second identity  $\vec{\nabla} \cdot (\vec{\nabla} \times \vec{A}) = 0$ , we start with the components of  $\vec{\nabla} \times \vec{A}$ , which are

$$(\nabla \times A)_{x} = \nabla_{y}A_{z} - \nabla_{z}A_{y}$$
$$(\nabla \times \vec{A})_{y} = \nabla_{z}A_{x} - \nabla_{x}A_{z}$$
$$(\nabla \times \vec{A})_{z} = \nabla_{x}A_{y} - \nabla_{y}A_{x}$$
(8)

Note that to get all three components of  $\vec{\nabla} \times \vec{A}$ , you do not have to memorize all three equations. If you memorize only the first  $(\vec{\nabla} \times \vec{A})_x = \nabla_y A_z - \nabla_z A_y$ you can get the other two by using *cyclic permutations*. That means, start with  $(\vec{\nabla} \times \vec{A})_x = \nabla_y A_z - \nabla_z A_y$ , and replace the subscripts cyclically, letting  $x \rightarrow y$ ,  $y \rightarrow z$ , and  $z \rightarrow x$ . That gives you  $(\vec{\nabla} \times \vec{A})_y = \nabla_z A_x - \nabla_x A_z$ . Do the cyclic permutation again and you get  $(\vec{\nabla} \times \vec{A})_z = \nabla_x A_y - \nabla_y A_x$  which is the third equation.)

Now take the dot product of  $\vec{\nabla}$  with  $\vec{\nabla} \times \vec{A}$  to get

$$\vec{\nabla} \cdot (\vec{\nabla} \times \vec{A}) = \nabla_{x} (\vec{\nabla} \times \vec{A})_{x} + \nabla_{y} (\vec{\nabla} \times \vec{A})_{y} + \nabla_{z} (\vec{\nabla} \times \vec{A})_{z}$$
$$= \nabla_{x} \nabla_{y} A_{z} - \nabla_{x} \nabla_{z} A_{y}$$
$$+ \nabla_{y} \nabla_{z} A_{x} - \nabla_{y} \nabla_{x} A_{z}$$
$$+ \nabla_{z} \nabla_{x} A_{y} - \nabla_{z} \nabla_{y} A_{x} \qquad (9)$$

#### Exercise 1

Show that all the terms in Equation (9) cancel, giving  $\nabla \cdot (\nabla \times \vec{A}) = 0$  for any  $\vec{A}$ .

## **Proof of Identity 3**

The third vector identity

$$\vec{\nabla} \times (\vec{\nabla} \times \vec{A}) = -\nabla^2 A + \vec{\nabla} (\vec{\nabla} \cdot \vec{A})$$
 (5) repeat

looks worse but is not that hard to prove. We will start with the  $\vec{x}$  component of  $\vec{\nabla} \times (\vec{\nabla} \times \vec{A})$  which is

$$\begin{split} & \left[ \vec{\nabla} \times (\vec{\nabla} \times \vec{A}) \right]_{x} = \nabla_{y} (\vec{\nabla} \times \vec{A})_{z} - \nabla_{z} (\vec{\nabla} \times \vec{A})_{y} \\ &= \nabla_{y} (\nabla_{x} A_{y} - \nabla_{y} A_{x}) - \nabla_{z} (\nabla_{z} A_{x} - \nabla_{x} A_{z}) \\ &= -\nabla_{y} \nabla_{y} A_{x} - \nabla_{z} \nabla_{z} A_{x} + \nabla_{x} \nabla_{y} A_{y} + \nabla_{x} \nabla_{z} A_{z} \end{split}$$

$$(10)$$

where we changed the order of differentiation in the last two terms. The trick is to add and then subtract  $\nabla_x \nabla_x A_x$  to Equation (10), giving

$$\begin{bmatrix} \vec{\nabla} \times (\vec{\nabla} \times \vec{A}) \end{bmatrix}_{x}$$

$$= -\nabla_{x} \nabla_{x} A_{x} - \nabla_{y} \nabla_{y} A_{x} - \nabla_{z} \nabla_{z} A_{x}$$

$$+ \nabla_{x} \nabla_{x} A_{x} + \nabla_{x} \nabla_{y} A_{y} + \nabla_{x} \nabla_{z} A_{z}$$

$$= -(\nabla_{x} \nabla_{x} + \nabla_{y} \nabla_{y} + \nabla_{z} \nabla_{z}) A_{x}$$

$$+ \nabla_{x} (\nabla_{x} A_{x} + \nabla_{y} A_{y} + \nabla_{z} A_{z})$$

$$= -\nabla^{2} A_{x} + \nabla_{x} (\vec{\nabla} \cdot \vec{A})$$
(11)

This is just the x component of Equation (5). Similar derivations verify the y and z components of that vector identity.

## DERIVATION OF THE WAVE EQUATION

We are now in a position to derive the wave equation for electromagnetic waves, starting from Maxwell's equations. We will use Maxwell's equations for empty space, because Maxwell's major discovery was that electric and magnetic fields could propagate through empty space in a wavelike manner, and that these waves were light waves.

Maxwell's equations in differential form are, from Equations (8-49) of Chapter 8

$$\vec{\nabla} \cdot \vec{E} = \frac{\rho}{\epsilon_0} \qquad Gauss' \, law$$

$$\vec{\nabla} \cdot \vec{B} = 0 \qquad no \, monopole$$

$$\vec{\nabla} \times \vec{B} = \mu_0 \vec{i} + \mu_0 \epsilon_0 \frac{\partial \vec{E}}{\partial t} \qquad Ampere's \, law$$

$$\vec{\nabla} \times \vec{E} = -\frac{\partial \vec{B}}{\partial t} \qquad Faraday's \, law$$

## Maxwell's Equations

\_ →

where  $\rho(x,y,z)$  is the electric charge density in coulombs per cubic meter, and  $\vec{i}(x,y,z)$  is the electric current density in amperes per square meter.

In empty space, where the charge density  $\rho(x,y,z)$ and the current density  $\vec{i}(x,y,z)$  are zero, we get

 $\vec{\nabla} \cdot \vec{E} = 0 \qquad Gauss' \, law \qquad (13a)$ 

 $\vec{\nabla} \cdot \vec{B} = 0$  no monopole (13b)

$$\vec{\nabla} \times \vec{B} = \mu_0 \epsilon_0 \frac{\partial \vec{E}}{\partial t}$$
 Ampere's law (13c)

$$\vec{\nabla} \times \vec{E} = -\frac{\partial B}{\partial t}$$
 Faraday's law (13d)

Maxwell's Equations in Empty Space

In our discussion of vector fields in the Physics text, we pointed out that a vector field is uniquely determined if we have general formulas for the volume and line integrals of that field. Now, working with differential equations, that statement becomes the rule that a vector field like  $\vec{E}$  is determined if we know the divergence  $\vec{\nabla} \cdot \vec{E}$  and the curl  $\vec{\nabla} \times \vec{E}$  at every point in space\*. There are four Maxwell equations because we have to specify both the divergence and the curl of both  $\vec{E}$  and  $\vec{B}$ .

Equation (10) tells us that in empty space, neither  $\vec{E}$ nor  $\vec{B}$  have a divergence  $(\vec{\nabla} \cdot \vec{E} = \vec{\nabla} \cdot \vec{B} = 0)$ , and we only have to deal with the curls of these fields.

The trick we use to get a wave equation from Equations (13) is to take the curl of Equations (13c) and (13d). This gives us

$$\vec{\nabla} \times (\vec{\nabla} \times \vec{B}) = \mu_0 \varepsilon_0 \vec{\nabla} \times \left(\frac{\partial \vec{E}}{\partial t}\right)$$
(14a)

$$\vec{\nabla} \times (\vec{\nabla} \times \vec{E}) = -\vec{\nabla} \times \left(\frac{\partial \vec{B}}{\partial t}\right)$$
 (14b)

where we took the constants  $\mu_0$  and  $\epsilon_0$  outside the derivative in Equation (14a).

\*(If we have a field known only in some region of space, like the velocity field of a fluid in a section of pipe, we can uniquely determine the field if we know the divergence and curl within that region, and also the normal components of the field at the region's surface.) The next step is to use the fact that we can interchange the order of partial differentiation to get

$$\vec{\nabla} \times \left[ \frac{\partial E(x,y,z,t)}{\partial t} \right] = \frac{\partial}{\partial t} \left[ \vec{\nabla} \times E(x,y,z,t) \right]$$
 (15)

and a similar result for  $\vec{\nabla} \times (\partial \vec{B} / \partial t)$  to give

$$\vec{\nabla} \times (\vec{\nabla} \times \vec{B}) = \mu_0 \varepsilon_0 \frac{\partial}{\partial t} (\vec{\nabla} \times \vec{E})$$
(16a)

$$\vec{\nabla} \times (\vec{\nabla} \times \vec{E}) = -\frac{\partial}{\partial t} (\vec{\nabla} \times \vec{B})$$
 (16b)

Notice that the right hand sides of Equations (16) involve  $(\vec{\nabla} \times \vec{E})$  and  $(\vec{\nabla} \times \vec{B})$  which are given by Maxwell's Equations (13c) and (13d) as

$$\vec{\nabla} \times \vec{E} = -\frac{\partial \vec{B}}{\partial t}$$
 (13d) repeated

$$\vec{\nabla} \times \vec{B} = \mu_0 \varepsilon_0 \frac{\partial \vec{E}}{\partial t}$$
 (13c) repeated

Thus Equations (16) can be written as

$$\vec{\nabla} \times (\vec{\nabla} \times \vec{B}) = \mu_0 \varepsilon_0 \frac{\partial}{\partial t} \left( -\frac{\partial \vec{B}}{\partial t} \right)$$
$$= -\mu_0 \varepsilon_0 \frac{\partial^2 B}{\partial t^2}$$
(17a)

$$\vec{\nabla} \times (\vec{\nabla} \times \vec{E}) = -\frac{\partial}{\partial t} \left( \mu_0 \epsilon_0 \frac{\partial \vec{E}}{\partial t} \right)$$
$$= -\mu_0 \epsilon_0 \frac{\partial^2 E}{\partial t^2}$$
(17b)

Notice that at this point  $\vec{E}$  and  $\vec{B}$  obey exactly the same differential equation.

The final step is to use the vector identity

$$\vec{\nabla} \times (\vec{\nabla} \times \vec{A}) = -\nabla^2 A + \vec{\nabla} (\vec{\nabla} \cdot \vec{A})$$
 (5) repeat

Since both  $\vec{\nabla}\cdot\vec{E}$  and  $\vec{\nabla}\cdot\vec{B}$  are zero in empty space, we have

$$\vec{\nabla} \times (\vec{\nabla} \times \vec{B}) = -\nabla^2 B \tag{18}$$

and the same for  $\vec{\nabla} \times (\vec{\nabla} \times \vec{E})$  to give us

$$-\nabla^{2}\vec{E} = -\mu_{0}\varepsilon_{0}\frac{\partial^{2}\vec{E}}{\partial t^{2}}$$
(19a)

$$-\nabla^{2}\vec{B} = -\mu_{0}\varepsilon_{0}\frac{\partial^{2}\vec{B}}{\partial t^{2}}$$
(19b)

Dividing through by  $\mu_0 \epsilon_0$  gives

$$\frac{1}{\mu_0 \varepsilon_0} \nabla^2 \vec{E} = \frac{\partial^2 \vec{E}}{\partial t^2}$$
(20a)

$$\frac{1}{\mu_0 \varepsilon_0} \nabla^2 \vec{B} = \frac{\partial^2 \vec{B}}{\partial t^2}$$
(20b)

## PLANE WAVE SOLUTION

Repeating Equations (20), we have

$$\frac{1}{\mu_0 \varepsilon_0} \nabla^2 \vec{E} = \frac{\partial^2 \vec{E}}{\partial t^2}$$
(20a)

$$\frac{1}{\mu_0 \varepsilon_0} \nabla^2 \vec{\mathbf{B}} = \frac{\partial^2 \vec{\mathbf{B}}}{\partial t^2}$$
(20b)

To interpret these equations, let us assume that  $\vec{E}$  and  $\vec{B}$  have the shape more or less like that shown in Figure (32-23b) reproduced here again. All we need from that picture is that both  $\vec{E}$  and  $\vec{B}$  vary only in the direction of motion (call this the  $\hat{x}$  direction) and in time. There is no change of  $\vec{E}$  and  $\vec{B}$  in the  $\hat{y}$  and  $\hat{z}$  directions. Such a wave is called a *plane wave*, because there are no variations within a plane.

Using the coordinate system added to Figure (32-23b), we see that  $\vec{E}$  is  $\hat{y}$  directed (we would call this  $\hat{y}$  polarized radiation) and  $\vec{B}$  is  $\hat{z}$  directed. The formulas for  $\vec{E}$  and  $\vec{B}$  can thus be written for this  $\hat{z}$  directed plane wave

$$\vec{E} = \hat{y}E_{v}(x,t) \tag{21a}$$

$$\vec{B} = \hat{z}B_{z}(x,t) \tag{21b}$$

where Equations (21a) and (21b) remind us that we are dealing with a plane wave with no x or y dependence.

As a result

$$\nabla_{y}\vec{E} = \hat{y}\frac{\partial E(x,t)}{\partial y} = 0$$

and the same for  $\nabla_z \vec{E}$ ,  $\nabla_v \vec{B}$  and  $\nabla_z \vec{B}$ . Thus

$$\nabla^{2}\vec{E} = (\nabla_{x}\nabla_{x}\vec{E} + \nabla_{y}\nabla_{y}\vec{E} + \nabla_{z}\nabla_{z}\vec{E})$$
  
$$= \nabla_{x}\nabla_{x}\vec{E} = \hat{y}\frac{\partial^{2}E_{y}}{\partial x^{2}}$$
(22a)

and

$$\nabla^2 \vec{B} = \hat{z} \frac{\partial^2 B_z}{\partial x^2}$$
(22b)

The time derivatives of the plane wave fields of Equations (21) are

$$\frac{\partial^2 \vec{E}}{\partial t^2} = \hat{y} \frac{\partial^2 E_y(x,t)}{\partial t^2}$$
(23a)

$$\frac{\partial^2 \vec{B}}{\partial t^2} = \hat{z} \frac{\partial^2 B_z(x,t)}{\partial t^2}$$
(23b)



b) Electric and magnetic fields produced by smoothly switching the antenna current.

When we use Equation (22a) for  $\nabla^2 \vec{E}$  and (23a) for  $\partial^2 \vec{E} / \partial t^2$  in Equation (20a), the unit vectors  $\hat{y}$  cancel and we are left with

$$\frac{1}{\mu_0 \varepsilon_0} \frac{\partial^2 E_y(x,t)}{\partial x^2} = \frac{\partial^2 E_y(x,t)}{\partial t^2}$$
(24a)

We get a similar equation for  $B_z$ , namely

$$\frac{1}{\mu_0 \varepsilon_0} \frac{\partial^2 B_z(x,t)}{\partial x^2} = \frac{\partial^2 B_z(x,t)}{\partial t^2}$$
(24b)

In our discussion of the one dimensional wave equation in Chapter 2 of this text we had as the formula for the wave equation

$$v_{wave}^{2} \frac{\partial^{2} y(x,t)}{\partial x^{2}} = \frac{\partial^{2} y(x,t)}{\partial t^{2}} \begin{array}{c} one \\ dimensional \\ wave \\ equation \end{array} (2-73)$$

Comparing this wave equation with Equation (24), we see that the plane wave of Figure (32-23b) obeys the one dimensional wave equation with

$$v_{wave}^{2} = \frac{1}{\mu_{0}\varepsilon_{0}}$$

$$v_{wave} = \frac{1}{\sqrt{\mu_{0}\varepsilon_{0}}}$$
(25)

From the wave equation alone we immediately find that the speed of the wave is  $1/\sqrt{\mu_0 \epsilon_0}$  which is the speed of light. We get this result without going through all the calculations we did in the Physics text to derive the speed of the electromagnetic pulse.

What we have shown in addition is that the speed of the wave does not depend on its shape. All we used was that E = E(x,t) without saying what the x dependence was. Thus both the series of pulses in Figure (32-23a) and the sinusoidal wave in (32-23b) should have the same speed  $1/\sqrt{\mu_0\epsilon_0}$ . This we were not able to show using the integral form of Maxwell's equations.

## THE THREE DIMENSIONAL WAVE EQUATION

We have seen that if  $\vec{E}$  and  $\vec{B}$  are plane waves, i.e., vector fields that vary in time and only one dimension, then Equations (20a) and (20b) become the one dimensional wave equation for  $\vec{E}$  and  $\vec{B}$ . Since Equations (20) do not single out any one direction as being special, we would get a wave equation for a plane wave moving in any direction, and we see that Equations (20) are three dimensional wave equations for waves traveling at a speed  $v_{wave}^2 = 1/\sqrt{\mu_0 \epsilon_0}$ . Rewriting these equations in terms of  $v_{wave}$  rather than  $\mu_0 \epsilon_0$  gives us the general form of the three dimensional wave equation

$$v_{wave}^2 \nabla^2 \vec{E} = \frac{\partial^2 E}{\partial t^2}$$
(26)

and the same for  $\vec{B}$ .

The form we will generally recognize as being the three dimensional wave equation is the trivial rearrangement of Equation (26),

$$\boxed{\frac{1}{v_{wave}^2} \frac{\partial^2 \vec{E}}{\partial t^2} - \nabla^2 \vec{E} = 0} \qquad three dimensional wave equation applied to \vec{E}}$$

(27)

Equation (27) is the way the wave equation is usually written in textbooks.

So far we have only shown that plane waves are a solution to the three dimensional wave equation. For now that is enough. Solutions to the wave equation can become quite complex in three dimensions, and we do not yet have to deal with these complications.

## APPENDIX: ORDER OF PARTIAL DIFFERENTIATION

It is worth while to show once and for all that you can interchange the order of partial differentiation. We do this by going back to the limiting process, where

$$\frac{\partial f(x,y)}{\partial x} = \lim_{\Delta x \to 0} \left[ \frac{f(x + \Delta x, y) - f(x,y)}{\Delta x} \right]$$
(A-1)

and a similar formula for  $\partial f/\partial y$ . For the second derivative we have

$$\nabla_{\mathbf{x}} \nabla_{\mathbf{y}} f(\mathbf{x}, \mathbf{y}) = \frac{\partial}{\partial \mathbf{x}} \left[ \frac{\partial f(\mathbf{x}, \mathbf{y})}{\partial \mathbf{y}} \right]$$
 (A-2)

Let us temporarily introduce the notation

$$f'_{y}(x,y) = \frac{\partial f(x,y)}{\partial y}$$
 (A-3)

so that Equation (A-2) becomes

$$\nabla_{x}\nabla_{y}f(x,y) = \frac{\partial}{\partial x}f'_{y}(x,y)$$
$$= \lim_{\Delta x \to 0} \left[\frac{f'_{y}(x+\Delta x,y) - f'_{y}(x,y)}{\Delta x}\right]$$
(A-4)

Now in Equation (A-4) make the substitution

$$\begin{aligned} f_{y}^{'}(x,y) &= \lim_{\Delta y \to 0} \left[ \frac{f(x,y+\Delta y) - f(x,y)}{\Delta y} \right] \text{ (A-5)} \\ f_{y}^{'}(x+\Delta x,y) &= \lim_{\Delta y \to 0} \left[ \frac{f(x+\Delta x,y+\Delta y) - f(x+\Delta x,y)}{\Delta y} \right] \end{aligned}$$

Using Equations (A-5) and (A-6) in (A-4) gives

$$\nabla_{x}\nabla_{y}f(x,y) = \lim_{\Delta x \to 0} \left[ \frac{f(x + \Delta x, y + \Delta y) + f(x,y) - f(x + \Delta x, y) - f(x, y + \Delta y)}{\Delta x \Delta y} \right]$$
(A-7)

## **Exercise 1**

Show that you get exactly the same result for  $\nabla_{y} \nabla_{x} f(x,y)$ .

You can see that our result, Equation (A-7) is completely symmetric between x and y, thus it should be obvious that we should get the same result by reversing the order of differentiation.

The only possible fly in the ointment is the order in which we take the limits as  $\Delta x \rightarrow 0$  and  $\Delta y \rightarrow 0$ . As long as f(x,y) is smooth enough so that f(x,y) and its first and second derivatives are continuous, then the order in which we take the limit makes no difference.
## Calculus 2000-Chapter 10 Conservation of Electric Charge

In this short chapter, we obtain a very important result. We will see that Maxwell's equations themselves imply that electric charge is conserved. In our development of Maxwell's equations, our attention was on the kind of electric and magnetic fields that were produced by electric charges and currents. We said, for example, that given some electric charge, Gauss' law would tell us what electric field it would produce. Or given an electric current, Ampere's law would tell us what magnetic field would result. Then later on, we found out that for mathematical consistency, a changing electric field would create a magnetic field and vice versa. All this was summarized in Maxwell's equations, which we repeat here

 $\vec{\nabla} \cdot \vec{E} = \rho/\epsilon_{0}$   $\vec{\nabla} \cdot \vec{B} = 0$   $\vec{\nabla} \times \vec{E} = -\partial B/\partial t$ (1)  $\vec{\nabla} \times \vec{B} = \mu_{0}\vec{i} + \mu_{0}\epsilon_{0}\partial\vec{E}/\partial t$ 

What we did not notice in this development of the equations for  $\vec{E}$  and  $\vec{B}$  is that the equations place a fundamental restriction on the sources  $\rho$  and  $\vec{i}$  of the fields. As we will now see, the restriction is that the electric charge, which is responsible for the charge density  $\rho$  and current  $\vec{i}$ , must be conserved.

### THE CONTINUITY EQUATION

We began our discussion of fluid dynamics in Chapter 23 of the Physics text, by introducing the continuity equation for an incompressible fluid. For a tube with an entrance cross sectional area  $A_1$  and exit area  $A_2$ , the equation was

$$v_1A_1 = v_2A_2$$
   
*continuity*  
*equation* (23-3)

which says that the same volume of fluid per second flowing into the entrance flows out of the exit. Later this statement that the fluid is incompressible (or does not get lost or created) became

$$\int_{\substack{\text{closed}\\\text{surface}}} \vec{v} \cdot d\vec{A} = 0 \qquad \begin{array}{c} \text{incompressible}\\ \text{fluid} \end{array}$$
(2)

The differential form of Equation (2) is

$$\vec{\nabla} \cdot \vec{v} = 0 \qquad \begin{array}{c} incompressible \\ fluid \end{array} \tag{3}$$

as we showed in our initial discussion of divergence. All three equations, (23-3), (2) and (3) are saying the same thing in a progressively more detailed way.

Equation (3) is not the most general statement of a continuity equation. It is the statement of the conservation of an incompressible fluid, but you can have flows of a compressible nature where something like mass or charge is still conserved. A more general form of the continuity equation allows for the conservation of these quantities. We will now see that this more general form of the continuity equations.

## CONTINUITY EQUATION FROM MAXWELL'S EQUATIONS

To derive the continuity equation for electric charge, we start by taking the divergence of the generalized form of Ampere's law

$$\vec{\nabla} \cdot \left[ \vec{\nabla} \times \vec{B} = \mu_0 \vec{i} + \mu_0 \varepsilon_0 \frac{\partial \vec{E}}{\partial t} \right]$$
(4)

which becomes

$$\vec{\nabla} \cdot (\vec{\nabla} \times \vec{B}) = \mu_0 \vec{\nabla} \cdot \vec{i} + \mu_0 \epsilon_0 \vec{\nabla} \cdot \left(\frac{\partial \vec{E}}{\partial t}\right)$$
(5)

Using the fact that the divergence of a curl is identically zero,  $\vec{\nabla} \cdot (\vec{\nabla} \times \vec{B}) = 0$ , and the fact that we can interchange the order of differentiation, we get

$$0 = \mu_0 \vec{\nabla} \cdot \vec{i} + \mu_0 \epsilon_0 \frac{\partial}{\partial t} (\vec{\nabla} \cdot \vec{E})$$
(6)

Divide Equation (6) through by  $\mu_0$ , and use Gauss' law

$$\vec{\nabla} \cdot \vec{E} = \frac{\rho}{\varepsilon_0}$$

to get

$$\vec{\nabla} \cdot \mathbf{i} + \varepsilon_0 \frac{\partial}{\partial t} \left( \frac{\rho}{\varepsilon_0} \right) = 0 \tag{7}$$

The  $\varepsilon_0$ 's cancel and we are left with

$$\frac{\partial \rho}{\partial t} + \vec{\nabla} \cdot \vec{i} = 0 \qquad \begin{array}{c} continuity equation \\ for electric charge \end{array} (8)$$

Equation (8) is the continuity equation for electric charge.

You can immediately see from Equation (8) that if the electric charge density  $\rho$  were unchanging in time, if  $\partial \rho / \partial t = 0$ , then we would have  $\vec{\nabla} \cdot i = 0$  and the electric current would flow as an incompressible fluid. The fact that a  $\partial \rho / \partial t$  term appears in Equation (8) is telling us what happens when  $\rho$  changes, for example, if we compress the charge into a smaller region.

#### Integral Form of Continuity Equation

The way to interpret Equation (8) is to convert the equation to its integral form. We do this by integrating the equation over some volume V bounded by a closed surface S. We have

$$\int_{V} \frac{\partial \rho}{\partial t} dV + \int_{V} \vec{\nabla} \cdot \vec{i} dV = 0$$
(9)

Using the divergence theorem to convert the volume integral of  $\vec{\nabla} \cdot \vec{i}$  to a surface integral gives

$$\int_{\substack{\text{volume} \\ V}} \vec{\nabla} \cdot \vec{i} \, dV = \int_{\substack{S \text{ (surface} \\ \text{ of } V)}} \vec{i} \cdot d\vec{A}$$
(10)

Using Equation (10) in (9) we get



(11)

On the left side of Equation (11) we have the term representing the net flow of electric current out through the surface S. It represents the total amount of electric charge per second leaving through the surface. On the right side we have an integral representing the rate at which the amount of charge remaining inside the volume V is decreasing (the – sign). Thus Equation (38) is telling us that the rate at which charge is flowing out through any closed surface S is equal to the rate at which the amount of charge remaining inside the surface is decreasing. This can be true for any surface S only if electric charge is everywhere conserved.

The fact that the continuity equation was a consequence of Maxwell's equation tells us that if we do have the correct equations for electric and magnetic fields, then the source of these fields, which is electric charge and current, must be a conserved source. Later, when we discuss the process of constructing theories of fields, we will see in more detail how conservation laws and theories of fields are closely related. Basically for every fundamental conservation law there is a field associated with the law. In this case the law is the conservation of electric charge and the associated field is the electromagnetic field. It turns out that the law of conservation of energy is associated with the gravitational field.

# Calculus 2000-Chapter 11 Scalar And Vector Potentials

In our first experiment on electricity in the Physics text we studied the relationship between voltage on electric fields. We constructed the lines of constant voltage, the equipotential lines, and then constructed the perpendicular electric field lines. In Chapter 3 of the Calculus text we developed the more detailed relationship that the electric field  $\vec{E}$  was equal to minus the gradient of the voltage

$$\vec{E}(x,y,z) = -\vec{\nabla}V(x,y,z) \tag{3-19}$$

As you study more advanced topics in science, you sometimes encounter situations where the name or symbol used to describe some quantity is different in the advanced texts than in the introductory ones. Various historical accidents are often responsible for this change.

In introductory texts and in the laboratory we talk about the voltage V which we measure with a voltmeter. The first hint that we would use a different name for voltage was when we called the lines of constant voltage **equipotential** lines, or lines of constant **potential**. Advanced texts, particularly those with a theoretical emphasis, use the name potential rather than voltage, and typically use the symbol  $\phi(x,y,z)$  rather than V(x,y,z). In this notation, Equation (3-19) becomes

$$\vec{E}(x,y,z) = -\vec{\nabla}\phi(x,y,z) \tag{1}$$

This is how we left the relationship between  $\vec{E}$  and  $\phi$  in Chapter 3 on gradients.

From our discussion of divergence and curl, it does not take long to see that there is a problem with Equation (1). If we take the curl of both sides of this equation, we get

$$\vec{\nabla} \times \vec{E} = -\vec{\nabla} \times (\vec{\nabla}\phi) \tag{2}$$

However our first vector identity, Equation (9-1) was that the curl of a divergence was identically zero.

$$\vec{\nabla} \times (\vec{\nabla}\phi) = 0 \tag{3}$$

Thus Equation (1) implies that the field  $\vec{E}$  has zero curl

$$\vec{V} \times \vec{E} = 0 \qquad \begin{array}{c} as \ a \ consequence \\ of \ Equation(1) \end{array}$$
(4)

which is not consistent with Maxwell's equations. In particular, Faraday's law says that

$$\vec{\nabla} \times \vec{E} = -\frac{\partial \vec{B}}{\partial t}$$
 Faraday's law (5)

Thus Equation (1) cannot be true, or at least cannot be the whole story, when changing magnetic fields are present, when  $\partial \vec{B}/\partial t$  is not zero. If we only have static charges, or even stationary currents so that  $\vec{B}$ is zero or constant in time, then Faraday's law becomes

$$\vec{V} \times \vec{E} = 0$$
  $\frac{when}{\partial \vec{B}/dt} = 0$  (6)

and then  $\vec{E}$  can be described completely as the gradient of a voltage V or potential  $\phi$ .

Since the curl is the line integral on an infinitesimal scale, Equation (6) is equivalent to the statement that the line integral of  $\vec{E}$  is zero everywhere

$$\oint \vec{E} \cdot d\vec{\ell} = 0 \qquad \stackrel{when}{\partial \vec{B}/dt} = 0 \qquad (6a)$$

In our initial discussion of the line integral in Chapter 28 of the Physics text (pages 28-5,6), we pointed out that Equation (6a) was the condition for what we called a **conservative force**, a force that could be described in terms of potential energy. The equation  $\vec{E} = -\vec{\nabla}\phi$  (or  $-\vec{\nabla}V$ ) does exactly that, since V or  $\phi$  is the potential energy of a unit test charge.

What we are seeing now is that for static fields, where  $\partial \vec{B}/\partial t$  is zero,  $\vec{E}$  is a conservative field that can be described as the gradient of a potential energy  $\phi$ . However when changing magnetic fields are present, the curl of  $\vec{E}$  is no longer zero and  $\vec{E}$ has a component that cannot be described as the gradient of a potential energy.

We will see in this chapter that  $\vec{E}$  and  $\vec{B}$  can both be described in terms of potentials by introducing a new kind of potential called the **vector potential**  $\vec{A}(x,y,z)$ . When combined with what we will now call the **scalar potential**  $\phi(x,y,z)$ , we not only have complete formulas for  $\vec{E}$  and  $\vec{B}$ , but also end up simplifying the electromagnetic wave equation for the case that sources like charge density  $\rho$  and current density  $\vec{i}$  are present.

The topic of the vector potential  $\vec{A}(x,y,z)$  is often left to later advanced physics courses, sometimes introduced at the graduate course level. There is no need to wait; the introduction of the vector potential provides good practice with curl and divergence. What we will not cover in this chapter are the ways the vector potential is used to solve complex radiation problems. That can wait. What we will focus on is how the vector potential can be used to simplify the structure of Maxwell's equations. In addition we need the vector potential to handle the concept of voltage when changing magnetic fields are present.

#### THE VECTOR POTENTIAL

It seems to be becoming a tradition in this text to begin each chapter with a repeat of Maxwell's equations. In order not to break the tradition, we do it again.

$$\vec{\nabla} \cdot \vec{E} = \frac{\rho}{\epsilon_0} \qquad Gauss' \, law$$

$$\vec{\nabla} \cdot \vec{B} = 0 \qquad no \, monopole$$

$$\vec{\nabla} \times \vec{B} = \mu_0 \vec{i} + \mu_0 \epsilon_0 \frac{\partial \vec{E}}{\partial t} \qquad Ampere's \, law$$

$$\vec{\nabla} \times \vec{E} = -\frac{\partial \vec{B}}{\partial t} \qquad Faraday's \, law$$
(7)

Let us now set the magnetic field  $\vec{B}(x,y,z)$  equal to the curl of some new vector field  $\vec{A}(x,y,z)$ . That is,

$$\vec{B}(x,y,z) \equiv \vec{\nabla} \times \vec{A}(x,y,z) \qquad introducing \\ the vector \\ potential \vec{A} \qquad (8)$$

Equation (7) is the beginning of our definition of what we will call the *vector potential*  $\vec{A}(x,y,z)$ . To begin to see why we introduced the vector potential, take the divergence of both sides of Equation (8). We get

$$\vec{\nabla} \cdot \vec{B} = \vec{\nabla} \cdot (\vec{\nabla} \times \vec{A}) = 0 \tag{9}$$

This is zero because of the second vector identity studied in Chapter 9, Equation (9-2). There we showed that the divergence of the curl  $\vec{\nabla} \cdot (\vec{\nabla} \times \vec{A})$  was identically zero for any vector field  $\vec{A}$ .

Thus if we define  $\vec{B}$  as the curl of some new vector field  $\vec{A}$ , then one of Maxwell's equations,  $\vec{\nabla} \cdot \vec{B} = 0$  is automatically satisfied.

Our next step is to see what happens when we introduce the vector potential into the other Maxwell equations. Let us start with Faraday's law

$$\vec{\nabla} \times \vec{E} = -\frac{\partial \vec{B}}{\partial t} \tag{10}$$

If we replace  $\vec{B}$  with  $\vec{\nabla} \times \vec{A}$  we get

$$\vec{\nabla} \times \vec{E} = -\frac{\partial}{\partial t} (\vec{\nabla} \times \vec{A}) \tag{11}$$

Using the fact that we can change the order of partial differentiation, and remembering that the curl is just a lot of partial derivatives, we get

$$\vec{\nabla} \times \vec{E} = \vec{\nabla} \times \left( -\frac{\partial \vec{A}}{\partial t} \right) \qquad Faraday's \ law \\ in \ terms \ of \ \vec{A} \qquad (12)$$

We see that Equation (12) would be satisfied if we could set  $\vec{E} = -\partial \vec{A} / \partial t$  on the left side.

We cannot do that, however, because we already know that for static charges,  $\vec{E} = -\vec{\nabla}\phi$ . But see what happens if we try the combination

$$\vec{E} = -\vec{\nabla}\phi - \frac{\partial\vec{A}}{\partial t}$$

$$electric field$$
in terms of
potentials
 $\phi$  and  $\vec{A}$ 

$$(13)$$

Taking the curl of Equation (7) gives

$$\vec{\nabla} \times \vec{E} = -\vec{\nabla} \times (\vec{\nabla}\phi) - \vec{\nabla} \times \frac{\partial \vec{A}}{\partial t}$$
(14)

Since  $\vec{\nabla} \times (\vec{\nabla} \phi) = 0$  because the curl of a gradient is identically zero, we get

$$\vec{\nabla} \times \vec{E} = -\vec{\nabla} \times \frac{\partial \vec{A}}{\partial t}$$
(15)

Next interchange the order of partial differentiation to get

$$\vec{\nabla} \times \vec{E} = -\frac{\partial}{\partial t} (\vec{\nabla} \times \vec{A}) = -\frac{\partial \vec{B}}{\partial t}$$
 (16)

which is Faraday's law.

Thus when we define the electric and magnetic fields  $\vec{E}$  and  $\vec{B}$  in terms of the potentials  $\phi$  and  $\vec{A}$  by

$$\vec{B} = \vec{\nabla} \times \vec{A}$$
 (8) repeated

$$\vec{E} = -\vec{\nabla}\phi - \partial\vec{A}/\partial t$$
 (13) repeated

then two of Maxwell's equations

$$\vec{\nabla} \cdot \vec{B} = 0$$
 no monopole

$$\vec{\nabla} \times \vec{E} = - \frac{\partial B}{\partial t}$$
 Faraday's law

are automatically satisfied.

You can now see how we handle potentials or voltages when changing magnetic fields are present. For the field of static charges, we have  $\vec{E} = -\vec{\nabla}\phi$  as before. When changing magnetic fields are present, we get an additional contribution to  $\vec{E}$  due to the  $-\partial\vec{A}/\partial t$  term.

In Maxwell's theory of electric and magnetic fields, in what is often called the *classical theory of electromagnetism*, you can solve all problems by using Maxwell's equations as shown in Equation (7) and never bother with introducing the vector potential  $\vec{A}$ . In the classical theory, the potentials are more of a mathematical convenience, trimming the number of Maxwell's equations from four to two because two of them are automatically handled by the definition of the potentials.

Things are different in quantum theory. There are experiments involving the wave nature of the electron that detect the vector potential  $\vec{A}$  directly. These experiments cannot be explained by the fields  $\vec{E}$  and  $\vec{B}$  alone. It turns out in quantum mechanics that the potentials  $\phi$  and  $\vec{A}$  are the fundamental quantities and  $\vec{E}$  and  $\vec{B}$  are derived concepts, concepts derived from the equations  $\vec{B} = \vec{\nabla} \times \vec{A}$  and  $\vec{E} = -\vec{\nabla} \phi - \partial \vec{A} / \partial t$ .

## WAVE EQUATIONS FOR $\phi$ and $\vec{\mathbf{A}}$

The other two Maxwell's equations turn out to be wave equations for  $\phi$  and  $\vec{A}$ . There is one surprise in store. So far we have defined only the curl of  $\vec{A}$ through the equation  $\vec{B} = \vec{\nabla} \times \vec{A}$ . In general a vector field like  $\vec{A}$  can have both a divergent part  $\vec{A}_{div}$  and a solenoidal part  $\vec{A}_{sol}$  where

$$\vec{A} = \vec{A}_{div} + \vec{A}_{sol} \tag{17}$$

where the divergent part has no curl and the solenoidal part has no divergence

$$\vec{\nabla} \times \vec{A}_{\text{div}} = 0 \tag{18a}$$

$$\vec{\nabla} \cdot \vec{A}_{sol} = 0 \tag{18b}$$

We saw this kind of separation in the case of electric fields. When the electric field was created by static electric charges it was purely divergent, i.e., had zero curl. An electric field created by a changing magnetic field is purely solenoidal, with zero divergence.

As a result our equation  $\vec{B} = \vec{\nabla} \times \vec{A}$  defines only the solenoidal part of  $\vec{A}$ , namely  $\vec{A}_{sol}$ . We are still free to choose  $\vec{A}_{div}$ , which has not been specified yet. We will see that we can choose  $\vec{A}_{div}$  or  $\vec{\nabla} \cdot \vec{A}$  in such a way that considerably simplifies the wave equations for  $\phi$  and  $\vec{A}$ . This choice is not essential, only convenient. Sometimes, in fact, it is more convenient not to specify any choice for  $\vec{A}_{div}$ , and to work with the more general but messier wave equations.

For very obscure historical reasons, the choice of a special value for  $\vec{\nabla} \cdot \vec{A}$  is called a *choice of gauge*. In a later chapter we will look very carefully at what it means to make different choices for  $\vec{\nabla} \cdot \vec{A}$ . We will see that there are no physical predictions affected in any way by changing our choice for  $\vec{\nabla} \cdot \vec{A}$ . As a result the theory of electromagnetism is said to be *invariant* under different choices of gauge, or *gauge invariant*. This feature of electromagnetism will turn out to have extremely important implications, particularly in the quantum theory. For now, however, we will simply make a special choice of  $\vec{\nabla} \cdot \vec{A}$  that simplifies the form of Maxwell's equations for  $\phi$  and  $\vec{A}$ .

The two Maxwell's equations that are not automatically satisfied by  $\vec{B} = \vec{\nabla} \times \vec{A}$  and  $\vec{E} = -\vec{\nabla}\phi - \partial \vec{A}/\partial t$ are

$$\vec{\nabla} \cdot \vec{E} = \frac{\rho}{\epsilon_0} \qquad Gauss' \, law$$
$$\vec{\nabla} \times \vec{B} = \mu_0 \vec{i} + \mu_0 \epsilon_0 \frac{\partial \vec{E}}{\partial t} \quad Ampere's \, law$$

Making the substitutions  $\vec{E} = -\vec{\nabla}\phi - \partial\vec{A}/\partial t$  in Gauss's law gives

$$\vec{\nabla} \cdot \vec{E} = \vec{\nabla} \cdot \left( -\vec{\nabla} \phi - \frac{\partial \vec{A}}{\partial t} \right) = \frac{\rho}{\epsilon_0}$$
(19)

Noting that  $\vec{\nabla} \cdot \partial \vec{A} / \partial t = \partial (\vec{\nabla} \cdot \vec{A}) / \partial t$  because we can change the order of partial differentiation, and that  $\vec{\nabla} \cdot (\vec{\nabla} \phi) = \nabla^2 \phi$ , we get

$$-\nabla^{2}\phi - \frac{\partial(\vec{\nabla}\cdot\vec{A})}{\partial t} = \frac{\rho}{\epsilon_{0}}$$
$$-\nabla^{2}\phi = \frac{\rho}{\epsilon_{0}} + \frac{\partial(\vec{\nabla}\cdot\vec{A})}{\partial t}$$
(20)

You can see the divergence of  $\vec{A}$ , namely  $\vec{\nabla} \cdot \vec{A}$  appearing in the equation for  $\phi$ .

Making the substitutions in Ampere's law gives

$$\vec{\nabla} \times \vec{B} = \vec{\nabla} \times (\vec{\nabla} \times \vec{A}) = \mu_0 \vec{i} + \mu_0 \varepsilon_0 \frac{\partial \vec{E}}{\partial t}$$
$$= \mu_0 \vec{i} + \mu_0 \varepsilon_0 \frac{\partial}{\partial t} \left( -\vec{\nabla} \phi - \frac{\partial \vec{A}}{\partial t} \right) \qquad (21)$$

Using the third vector identity of Chapter 9, namely

$$\vec{\nabla} \times (\vec{\nabla} \times \vec{A}) = -\nabla^2 \vec{A} + \vec{\nabla} (\vec{\nabla} \cdot \vec{A})$$
(9-3)

Equation 21 becomes

$$-\nabla^{2}\vec{A} + \vec{\nabla}(\vec{\nabla}\cdot\vec{A})$$

$$= \mu_{0}\vec{i} - \mu_{0}\epsilon_{0}\frac{\partial(\vec{\nabla}\phi)}{\partial t} - \mu_{0}\epsilon_{0}\frac{\partial^{2}\vec{A}}{\partial t^{2}}$$
(22)

Writing  $\partial(\vec{\nabla}\phi)/\partial t = \vec{\nabla}(\partial\phi/\partial t)$  and moving the  $\partial^2(\vec{A})/\partial t^2$  term to the left and  $\vec{\nabla}(\vec{\nabla}\cdot\vec{A})$  to the right gives

$$-\nabla^{2}\vec{A} + \mu_{0}\varepsilon_{0} \frac{\partial^{2}(A)}{\partial t^{2}}$$
$$= \mu_{0}\vec{i} - \vec{\nabla} \left(\mu_{0}\varepsilon_{0}\frac{\partial\phi}{\partial t}\right) - \vec{\nabla}(\vec{\nabla}\cdot\vec{A})$$
(23)

In Equation (23) we see the wave equation for  $\vec{A}$  appearing on the left side, but we have some weird stuff involving  $\vec{\nabla} \cdot \vec{A}$  and  $\partial \phi / \partial t$  on the right. We can simplify things a bit by noting that both of these terms have a factor of  $\vec{\nabla}$  and writing

where we have replaced  $\mu_0 \epsilon_0$  by  $1/c^2$ , c being the speed of light.

Equation (24) is beginning to look like a wave equation with some peculiar terms on the right hand side. Equation (20) for  $\phi$  does not, at least now, look like a wave equation. However we can make it look like a wave equation by adding the term  $(1/c^2)(\partial^2 \phi/\partial t^2)$  to both sides, giving

$$-\nabla^{2}\phi + \frac{1}{c^{2}}\frac{\partial^{2}(\phi)}{\partial t^{2}}$$
$$= \frac{\rho}{\varepsilon_{0}} + \frac{1}{c^{2}}\frac{\partial^{2}(\phi)}{\partial t^{2}} + \frac{\partial(\vec{\nabla}\cdot\vec{A})}{\partial t}$$
(25)

We can factor out a  $\partial/\partial t$  in the last two terms on the right side of Equation (25) giving us

$$-\nabla^{2}\phi + \frac{1}{c^{2}}\frac{\partial^{2}(\phi)}{\partial t^{2}}$$

$$= \frac{\rho}{\varepsilon_{0}} + \frac{\partial}{\partial t} \left( \vec{\nabla} \cdot \vec{A} + \frac{1}{c^{2}}\frac{\partial \phi}{\partial t} \right)$$
*Gauss' law*
(26)

The rather messy looking Equations (24) and (26) are Ampere's law and Gauss' law written in terms of the scalar and vector potentials  $\phi$  and  $\vec{A}$ .

On the left side of each we have the beginning of a wave equation, but somewhat of a mess on the right. However we see that the term

$$\vec{\nabla} \cdot \vec{A} + \frac{1}{c^2} \frac{\partial \phi}{\partial t}$$
(27)

is common to both equations. If we could find some way to get rid of this term, there would be a considerable simplification.

We have, however, not yet specified what the value of  $\vec{\nabla} \cdot \vec{A}$  should be. We have only specified  $\vec{\nabla} \times \vec{A} = \vec{B}$ . If we make the choice

$$\vec{\nabla} \cdot \vec{A} = -\frac{1}{c^2} \frac{\partial \phi}{\partial t} \quad \begin{array}{c} special \\ choice \\ of \ gauge \end{array}$$
(28)

then the term (27) goes to zero. Making a choice for  $\vec{\nabla} \cdot \vec{A}$  is called *making a choice of gauge*, and this particular choice leads to the much simpler equations

$$-\nabla^2 \phi + \frac{1}{c^2} \frac{\partial^2 \phi}{\partial t^2} = \frac{\rho}{\varepsilon_0} \qquad Gauss' law \qquad (29)$$

$$-\nabla^{2}\vec{A} + \frac{1}{c^{2}}\frac{\partial^{2}\vec{A}}{\partial t^{2}} = \mu_{0}\vec{i} \quad Ampere's \ law \quad (30)$$

We get the rather elegant result that both potentials, the scalar potential  $\phi$  and vector potential  $\vec{A}$ , obey wave equations with source terms on the right hand side. The source for the scalar potential is the charge density  $\rho/\epsilon_0$ , and the source for the vector potential is the current density  $\mu_0 \vec{i}$ .

#### **Exercise 1**

The choice of gauge we made to get Equations (29) and (30) was  $\vec{\nabla} \cdot \vec{A} = -(1/c^2)\partial\phi/\partial t$ . This gave us simple wave equations which are convenient if we are working with electromagnetic waves. Sometimes another choice of gauge is more convenient. Derive Gauss' law and Ampere's law in terms of  $\phi$  and  $\vec{A}$ , using the choice of gauge

$$\vec{\nabla} \cdot \vec{A} = 0$$
 Coulomb gauge (31)

which is called the *Coulomb gauge*.

Do this derivation two ways. One by starting from Maxwell's equations in terms of  $\vec{E}$  and  $\vec{B}$ , and secondly, starting from Equations (24) and (26) where we made no special choice of gauge.

#### Exercise 2

This exercise is optional, but should give some very good practice with Maxwell's equations. In Chapter 9 we derived the wave equation for electromagnetic waves in empty space by first writing Maxwell's equations for empty space, Equations (9-12), and then taking the curl of Ampere's and Faraday's law. The results were

$$-\nabla^{2}\vec{\mathsf{E}} + \frac{1}{c^{2}}\frac{\partial^{2}\vec{\mathsf{E}}}{\partial t^{2}} = 0$$
  
$$-\nabla^{2}\vec{\mathsf{B}} + \frac{1}{c^{2}}\frac{\partial^{2}\vec{\mathsf{B}}}{\partial t^{2}} = 0 \qquad \text{wave equations} \\ \text{in empty space} \qquad (9-20)$$

Now repeat these calculations for the case that the charge and current densities  $\rho$  and i are not zero. Show that you get the following wave equations for  $\vec{E}$  and  $\vec{B}$ 

$$-\nabla^{2}\vec{\mathsf{E}} + \frac{1}{c^{2}}\frac{\partial^{2}\vec{\mathsf{E}}}{\partial t^{2}} = -\frac{\vec{\nabla}\rho}{\varepsilon_{0}} - \mu_{0}\frac{\partial\vec{\mathsf{i}}}{\partial t}$$
(32)

$$-\nabla^{2}\vec{B} + \frac{1}{c^{2}}\frac{\partial^{2}\vec{B}}{\partial t^{2}} = \mu_{0}\vec{\nabla}\times\vec{i}$$
(33)

You can see that we still get wave equations for  $\vec{E}$  and  $\vec{B}$ , but the source terms, the stuff on the right hand side, are much more complex than the source terms for the wave equations for  $\phi$  and  $\vec{A}$ . For example, the source term for the  $\vec{A}$  wave is simply  $\mu_0 \vec{i}$ , while the source term for a  $\vec{B}$  wave is the  $\mu_0 \vec{\nabla} \times \vec{i}$ . It is even worse for the  $\vec{E}$  field. Instead of the source term  $\rho/\epsilon_0$  for the  $\phi$  field, we have  $(-\nabla \rho/\epsilon_0 - \mu_0 \partial \vec{i}/\partial t)$  as a source for the  $\vec{E}$  wave.

#### Summary

Here we collect in one place, all the forms of Maxwell's equations.

(a) Maxwell's equations in terms of  $\vec{E}$  and  $\vec{B}$ 

$$\vec{\nabla} \cdot \vec{E} = \frac{\rho}{\epsilon_0} \qquad Gauss' \, law$$
$$\vec{\nabla} \cdot \vec{B} = 0 \qquad no \, monopole$$
$$\vec{\nabla} \times \vec{B} = \mu_0 \vec{i} + \mu_0 \epsilon_0 \frac{\partial \vec{E}}{\partial t} \qquad Ampere's \, law$$
$$\vec{d} = \vec{D} \vec{B}$$

$$\vec{\nabla} \times \vec{E} = -\frac{\partial B}{\partial t}$$
 Faraday's law

(b) Wave equations for  $\vec{E}$  and  $\vec{B}$ 

$$\begin{split} &-\nabla^2 \vec{E} + \frac{1}{c^2} \frac{\partial^2 \vec{E}}{\partial t^2} = -\frac{\vec{\nabla} \rho}{\epsilon_0} - \mu_0 \frac{\partial \vec{i}}{\partial t} \\ &-\nabla^2 \vec{B} + \frac{1}{c^2} \frac{\partial^2 \vec{B}}{\partial t^2} = \mu_0 \vec{\nabla} \times \vec{i} \end{split}$$

For the wave equations in empty space, set  $\rho = 0$  and  $\vec{i} = 0$ .

(c) Scalar and vector potentials  $\phi$  and  $\vec{A}$ 

$$\vec{\mathbf{B}} = \vec{\nabla} \times \vec{\mathbf{A}}$$
$$\vec{\mathbf{E}} = -\vec{\nabla} \phi - \partial \vec{\mathbf{A}} / \partial t$$

These automatically satisfy

$$\vec{\nabla} \cdot \vec{B} = 0$$
$$\vec{\nabla} \times \vec{E} = -\partial \vec{B} / \partial t$$

The remaining two Maxwell's equations become

$$-\nabla^{2}\phi + \frac{1}{c^{2}}\frac{\partial^{2}\phi}{\partial t^{2}} = \frac{\rho}{\varepsilon_{0}} + \frac{\partial}{\partial t}\left[\vec{\nabla}\cdot\vec{A} + \frac{1}{c^{2}}\frac{\partial\phi}{\partial t}\right]$$
$$-\nabla^{2}\vec{A} + \frac{1}{c^{2}}\frac{\partial^{2}\vec{A}}{\partial t^{2}} = \mu_{0}\vec{i} - \vec{\nabla}\left[\vec{\nabla}\cdot\vec{A} + \frac{1}{c^{2}}\frac{\partial\phi}{\partial t}\right]$$

The terms in the square brackets can be set to zero with the choice of gauge

$$\vec{\nabla} \cdot \vec{A} = -\frac{1}{c^2} \frac{\partial \phi}{\partial t}$$
 special choice  
of  $\vec{\nabla} \cdot \vec{A}$ 

With this choice of gauge, Maxwell's equations reduce to

$$\begin{split} &-\nabla^2 \varphi + \frac{1}{c^2} \frac{\partial^2 \varphi}{\partial t^2} \, = \, \frac{\rho}{\epsilon_0} \\ &-\nabla^2 \vec{A} + \frac{1}{c^2} \frac{\partial^2 \vec{A}}{\partial t^2} \, = \, \mu_0 \vec{i} \end{split}$$

all that is left of Maxwell's equations

## Calculus 2000-Chapter 12 Vorticity

At the beginning of Part II of the Physics text, we used the velocity field to introduce the concept of a vector field. It is easier to picture velocity vectors attached to water molecules in a flowing stream than to visualize a vector at each point in space. We could introduce Gauss' law as a conservation law for an incompressible fluid, and then show that the electric field behaved in a similar way.

Since that early introduction, we have come a long way in our study of the mathematical behavior of vector fields. In this and the next chapter, we will turn the tables on our earlier approach and apply to the velocity field the techniques and insights we have gained in our study of electric and magnetic fields. This will lead to a much deeper understanding of the behavior of fluids than we got in our old discussion of Bernoulli's equation. The most important concept that carries us beyond Bernoulli's equation is **vorticity**, which is the curl of the velocity field. Vorticity is important not only in the study of vortex structures like vortex rings and tornadoes, it plays a fundamental role in all aspects of fluid motion. In this chapter, we will develop an intuitive picture of vorticity. In the next chapter, we focus on its dynamic behavior.

These two chapters are designed to be an introduction to the basic concepts of **fluid dynamics**. For most of the past century, this subject has been eliminated from the undergraduate physics curriculum, despite exciting advances in the understanding of the behavior of superfluids. One of our aims with these chapters is to bring this subject back.

### **DIVERGENCE FREE FIELDS**

In the Physics text, we have often noted the similarity between the magnetic field and the velocity field. The fact that there are no magnetic charges led to the equation

$$\int_{S} \vec{B} \cdot d\vec{A} = 0 \qquad \begin{array}{c} \text{for any closed} \\ \text{surface } S \end{array}$$
(1)

For an incompressible fluid like water, the continuity equation, i.e., the fact that we cannot create or destroy water molecules, leads to the equation

$$\int_{S} \vec{v} \cdot d\vec{A} = 0 \qquad for any closed \\ surface S \qquad (2)$$

With the introduction of our differential notation, we saw that Equation (1) for the magnetic field became

$$\vec{\nabla} \cdot \vec{B} = 0 \tag{1a}$$

The same mathematics leads to the equation for the velocity field

$$\vec{\nabla} \cdot \vec{v} = 0 \qquad \begin{array}{c} continuity equation \\ for an \\ incompressible fluid \end{array} (2a)$$

Thus we see that both the magnetic field, and the velocity field of an incompressible fluid, are divergence free fields.

Another way to see the same result is to look at the form of the continuity equation we discussed a short while ago in Chapter 10. We saw how Maxwell's equations automatically led to a continuity equation for electric charge. That equation was

$$\frac{\partial \rho}{\partial t} + \vec{\nabla} \cdot \vec{i} = 0 \quad \begin{array}{c} \text{continuity equation} \\ \text{for electric charge} \end{array} \quad (Cal \ 10-8)$$

When applied to a fluid of mass density  $\rho$  and mass current density  $\rho \vec{v}$  the continuity equation for mass becomes

$$\frac{\partial \rho}{\partial t} + \vec{\nabla} \cdot (\rho \vec{v}) = 0 \qquad \begin{array}{c} \text{continuity equation} \\ \text{for a fluid of} \\ \text{mass density } \rho \end{array} (3)$$

If the fluid density  $\rho$  is constant, then  $\partial \rho / \partial t = 0$  and  $\vec{\nabla} \rho = 0$ . This leads to  $\vec{\nabla} \cdot (\rho \vec{v}) = \rho \vec{\nabla} \cdot \vec{v} = 0$  and we are left with

$$\vec{\nabla} \cdot \vec{v} = 0$$
 (2a) repeated

as the continuity equation for a constant density fluid.

## THE VORTICITY FIELD

When we were discussing electric and magnetic fields in the Physics text, we found that we needed equations for both the surface integral and the line integral in order to specify the field. That is why we ended up with four Maxwell's equations in order to describe the two fields  $\vec{E}$  and  $\vec{B}$ . In the Calculus text, we have shrunk the surface and line integrals down to infinitesimal size where they become the divergence and the curl. Thus to specify a field, we now need equations for both the divergence and curl of the field.

As we mentioned in Chapter 9, if we have a field known only in some limited volume of space, like the velocity field of a fluid within a section of pipe, then *in order to uniquely determine the field, we must know not only the divergence and curl within that volume, but also the perpendicular components of the field at the volume's surface*. It is the perpendicular components of the velocity field at the volume's surface that tell us how the fluid is flowing in and out.

For a constant density or incompressible fluid, we already know that the divergence is zero. Thus if we know how the fluid is flowing into and out of a volume, the only other thing we need to specify is its curl  $\vec{\nabla} \times \vec{v}$  inside. From this point of view we see that the curl  $\vec{\nabla} \times \vec{v}$  plays a key role in determining the nature of fluid flows. It should thus not be too surprising that most of this chapter is devoted to understanding the nature and behavior of the curl  $\vec{\nabla} \times \vec{v}$ .

Our first step will be to give the curl  $\vec{\nabla} \times \vec{v}$  a name. We will call it *vorticity* and designate it by the Greek letter  $\vec{\omega}$  (omega).

$$\vec{\boldsymbol{\omega}} \equiv \vec{\nabla} \times \vec{v} \qquad \text{vorticity} \qquad (4)$$

At this point, we have a slight problem with notation. In the Physics text we used the symbol  $\omega$  to designate angular velocity  $d\theta/dt$ . While there is some relationship between angular velocity  $d\theta/dt$  and vorticity  $\vec{\omega} = \vec{\nabla} \times \vec{v}$ , they are different quantities. Worse yet, in one important example, namely the rotation of a solid body, they differ by exactly a factor of 2. To avoid ambiguity, we will in this chapter use  $\vec{\omega}$  for vorticity  $\vec{\nabla} \times \vec{v}$ , and the symbol  $\omega_{rot}$  for angular velocity.

$$\omega_{\rm rot} \equiv \frac{\mathrm{d}\theta}{\mathrm{d}t} \qquad angular \, velocity \qquad (5)$$

#### **POTENTIAL FLOW**

In the next few sections, we will develop an intuition for the concept of vorticity by considering various examples. We will start with the simplest example, namely flow with no vorticity, i.e., when  $\vec{\nabla} \times \vec{v} = 0$ . Such flows are called *potential flows*. The reason for the name is as follows.

In our early discussion of electric fields, we pointed out that both the gravitational field, and the electric field of stationary point charges were conservative fields. A conservative field was defined as one where the total work done by the field acting on a mass or charge was zero if we carried the particle around and came back to the original starting point. (See page 25-5 of the Physics text.) For the work done by an electric field on a unit test charge, this statement took the form

$$\oint \vec{E} \cdot d\vec{\ell} = 0 \qquad \begin{array}{c} \text{condition that } \vec{E} \\ \text{is a conservative field} \end{array} \tag{6}$$

In our differential notation, Equation (6) becomes

$$\vec{\nabla} \times \vec{E} = 0$$
 condition that  $\vec{E}$  (7)  
is a conservative field

You will recall that when  $\vec{E}$  was a conservative field, we could introduce a unique potential energy provided we defined the zero of potential energy. We called the potential energy of a unit test charge *electric voltage* or *electric potential*.

When we got to Faraday's law, we had some problems with the concept of electric voltage. In our discussion of the betatron where electrons are circling a region of changing magnetic flux, the electrons gained voltage each time they went around the circle. When a changing magnetic field or magnetic flux  $\Phi_B$  is present, the voltage or electric potential is not unique because the electric field is no longer a conservative field. Faraday's law in integral and differential form is

$$\oint \vec{E} \cdot d\vec{\ell} = -\frac{d\Phi_{\rm B}}{dt} \qquad (\text{Physics 32-19})$$

$$\vec{\nabla} \times \vec{E} = -\frac{dB}{dt} \tag{8-49}$$

and we see that  $\vec{\nabla} \times \vec{E}$  is no longer zero.

When  $\vec{\nabla} \times \vec{E}$  is zero we have a unique electric voltage (once we have defined the zero of voltage), and we can use the concept of the gradient, discussed in the Calculus Chapter 3, to calculate the electric field from the voltage. The formula we had was

$$\vec{E} = -\vec{\nabla}V(x, y, z)$$
(3-19)

where V(x, y, z) is the voltage.

By similar arguments, if we have a conservative velocity field  $\vec{v}$ , one obeying the condition

$$\vec{\nabla} \times \vec{v} = 0$$
 conservative velocity field (8)

then we can introduce potential  $\varphi(x, y, z)$  that is analogous to the voltage V(x, y, z) for the electric field. In terms of the potential  $\varphi$ , the velocity field  $\vec{v}$  would be given by

$$\vec{v} = -\vec{\nabla}\phi$$
 velocity field derived  
from a potential (9)

Because such a velocity field is derived from a potential  $\varphi$ , the flow field is called *potential flow*.

As a quick check that our formulas are working correctly, suppose we start with some potential flow  $\vec{v} = -\vec{\nabla}\phi$  and ask what its curl is. We have

$$\vec{\nabla} \times \vec{v} = \vec{\nabla} \times (-\nabla \phi) \tag{10}$$

One of the vector identities, from Calculus Chapter 9 was

$$\vec{\nabla} \times (\vec{\nabla} f) = 0 \tag{9-1}$$

where f is any scalar function. Thus  $\vec{\nabla} \times (\vec{\nabla} \phi)$  is identically zero, and any flow derived from a potential  $\phi$  has to have zero curl, or no vorticity.

#### **Examples of Potential Flow**

If we combine the equation  $\vec{v} = -\vec{\nabla}\phi$  for potential flow with the divergence free condition  $\vec{\nabla} \cdot \vec{v} = 0$  we get

 $\vec{\nabla} \cdot \vec{v} = \vec{\nabla} \cdot (-\vec{\nabla}\phi) = 0$ or  $\nabla^2 \phi = 0$ (11)

The operator  $\nabla^2$  is the Laplacian operator we discussed in detail in Chapter 4. Equation (11) itself is known as *Poisson's equation*.

To find examples of potential flow, one can use Equation (11) subject to the boundary conditions on the velocity field at the walls of the container. A number of techniques have been developed to solve this problem, both approximation techniques for analytical solutions and numerical techniques for computer solutions. We are not going to discuss these techniques because the work is hard and the results are not particularly applicable to real fluid flows. We will see that almost all fluid flows involve vorticity, and our interest in this chapter will be the behavior of the vorticity. When we need a potential flow solution, we will either choose one simple enough to guess the shape or rely on someone else's solution.

#### **Potential Flow in a Sealed Container**

As our first example, suppose we have a constant density fluid in a completely sealed container. That means that no fluid is flowing in or out. Now suppose the fluid has no vorticity, that  $\vec{\nabla} \times \vec{v} = 0$  inside. The resulting flow then must be a potential flow.

One possible solution for  $\vec{\nabla} \times \vec{v} = 0$  is that the fluid inside is at rest (assuming that the container walls are at rest). That is,

$$\vec{v} = 0$$
 a potential flow  
solution for a (12)  
sealed container

This solution clearly obeys the condition  $\vec{\nabla} \times \vec{v} = 0$ and  $\vec{\nabla} \cdot \vec{v} = 0$ , and has no normal flow at the boundary walls.

What other potential flow solutions are there? **NONE**. Our mathematical theorem given at the beginning of the chapter states that the vector field  $\vec{v}$  is uniquely determined if we specify  $\vec{\nabla} \cdot \vec{v}$  and  $\vec{\nabla} \times \vec{v}$  within a closed volume V and the normal components of  $\vec{v}$  at the surface of V. We have done that. Thus the solution  $\vec{v} = 0$  is unique, and there is no other potential flow solution.

This solution emphasizes the importance of vorticity in the study of fluid flows. If we have a sealed container filled with a constant density fluid, there can be no flow without vorticity. *In this case, the source of all fluid motion must be vorticity.* This is why it is so important in the study of fluid behavior to understand the role and behavior of vorticity.

#### **Potential Flow in a Straight Pipe**

We began our discussion of fluid motion in Chapter 23 of the Physics text, with the example of a fluid entering a pipe at a velocity  $\vec{v}_1$  and exiting at a velocity  $\vec{v}_2$  as shown in Figure (1). We assumed that  $\vec{v}_1$  was uniform over the entire inlet and  $\vec{v}_2$  over the entire exit. The continuity equation gave  $v_1A_1 = v_2A_2$ . If the pipe is uniform, so that  $A_1 = A_2$ , we get  $v_1 = v_2$ .

What is the potential flow solution for the uniform pipe of Figure (1)? One possible answer is shown in Figure (2), namely that the velocity field is a constant throughout the pipe.

$$\vec{v} = \vec{v}_1 = \text{constant} \quad \begin{array}{c} \text{potential} \\ \text{flow solution} \end{array}$$
(13)

Let us check that  $\vec{v} = \vec{v}_1 = \text{constant}$  is a potential flow solution. It is clear that the divergence  $\vec{\nabla} \cdot \vec{v}_1$ and the curl  $\vec{\nabla} \times \vec{v}_1$  are both zero for a constant vector field  $\vec{v}_1$ . Thus the flow  $\vec{v} = \vec{v}_1$  is potential flow. The solution  $\vec{v} = \vec{v}_1$  also has the correct normal components, being  $\vec{v}_1$  at the entrance and exit, and no normal flow at the pipe walls. Thus Figure (2), with  $\vec{v} = \vec{v}_1 = \text{constant}$ , is our unique solution for potential flow in a straight pipe with uniform entrance and exit velocities. As we said, in some cases we can guess the potential flow solutions. The problem with the potential flow solution of Figure (2) is that a fluid like water cannot flow that way. In Figure (2), the fluid is slipping at the pipe walls. The first layer of atoms next to the walls is moving just as fast as the atoms in the center of the flow. For all normal fluids the first layer of atoms is stuck to the wall by molecular forces, and due to viscous effects, the fluid velocity has to increase gradually as we go into the fluid. There is no potential flow solution for pipe flow that has this property, thus all flows of normal fluids in a pipe must involve vorticity.



#### Figure 1

A fluid enters a uniform pipe at a velocity  $\vec{v}_1$ .



#### Figure 2

One possible solution to the potential flow problem. If we have a uniform pipe, with a uniform inlet and outflow velocities as shown in Figure (1), then this is the only solution.

## SUPERFLUIDS

Normal fluids like water cannot slip along the surface of a pipe, but superfluids, which have zero viscosity, can. As a result a superfluid can have a potential flow pattern like that shown in Figure (2). We have good experimental evidence that in a number of examples superfluid helium does flow that way.

In the 1940s, the Russian physicist Lev Landau made the prediction, based on his wave equation for the atoms in a superfluid, that superfluid helium had to flow without vorticity, that  $\vec{\nabla} \times \vec{v} = 0$  and only potential flow solutions would be possible. This was a prediction that was fairly easy to check by the following experiment.

If you place a glass of water on a spinning turntable and wait until the water rotates with the glass, the surface of the water will be slightly curved, as the water is pushed to the outside by "centrifugal forces". (If you choose a coordinate system that is rotating with the glass, then in this rotating coordinate system there is an outward centrifugal pseudo force.) The shape of the surface of the water turns out to be a parabola. In fact, large modern telescopes are now made by cooling the molten glass in a rotating container so that the rough parabolic shape is already there when the glass hardens.

Now consider how superfluid helium should behave when in such a rotating container. If the container is circular, like a drinking glass, and centered on the axis of rotation, the container can rotate without forcing the fluid to have any sideways motion. Also no fluid is flowing into or out of the bottom or top. Thus the normal or perpendicular component of flow is zero all around the fluid.

Superfluid helium is essentially a constant density fluid, thus  $\vec{\nabla} \cdot \vec{v} = 0$  within the fluid. If Landau were right, then  $\vec{\nabla} \times \vec{v}$  should also be zero inside the fluid, and we would have to have potential flow.

We have already discussed the potential flow solution for this case. If there is no normal flow through the fixed boundaries of the fluid, the unique potential flow solution for a constant density fluid is  $\vec{v} = 0$ . The fluid cannot rotate with the bucket. It cannot move at all! We get the unique prediction that the fluid must be at rest, and as a result the surface of the fluid must be flat. This prediction is easy to test; rotate a bucket of superfluid helium and see if the surface is flat or parabolic.

There are a few complications to the experiment. Above a temperature of 2.17 kelvins, liquid helium is a normal fluid with viscosity like other fluids with which we are familiar. When helium is cooled to just below 2.17 kelvins, superfluidity sets in, but in a rather peculiar way. The best way to understand the properties of liquid helium below 2.17 k is to think of it as a mixture of two fluids, a normal fluid with viscosity and a superfluid with no viscosity. At the temperature 2.17 k, the helium is almost all normal fluid. As we cool further, we get more superfluid and less normal fluid. Down at a temperature of 1 kelvin, which is quite easy to reach experimentally, almost all the normal fluid is gone and we have essentially pure superfluid.

In Landau's picture, the normal fluid below 2.17 k has viscosity, is not bound by the condition  $\vec{\nabla} \times \vec{v} = 0$ , and thus can rotate. Only the superfluid component must have  $\vec{\nabla} \times \vec{v} = 0$  and undergo only potential flow. Thus if we have a rotating bucket of superfluid helium at just below 2.17 k, it should be mostly normal fluid and eventually start rotating with the bucket. We should expect to see a parabolic surface, and that is what is seen experimentally.

However, as we cool the helium from just below 2.17 k down to 1 k, the normal fluid turns to superfluid. If Landau were right, the flow should go over to a potential flow and the surface of the liquid should become flat even though the container keeps rotating. This does not happen, and something has to be wrong with Landau's prediction. The curved surface at 1 k indicates that the superfluid is moving, and thus must contain some vorticity. In a later section we will see how Feynman was able to explain the parabolic surface, while still obeying Landau's condition  $\vec{\nabla} \times \vec{v} = 0$  almost everywhere in the fluid.

## **VORTICITY AS A SOURCE OF FLUID MOTION**

In our discussion of potential flow of a constant density fluid in a sealed container, we saw that there could be no flow without vorticity. Vorticity must be the source of any flow found there. In this section, we will illustrate the idea that vorticity is the source of fluid motion by comparing the velocity field with the magnetic field of electric currents. We will see that vorticity is a source of the velocity field in much the same way that an electric current is a source of the magnetic field.

In our discussion of magnetic fields, it was clear that magnetic fields are created by electric currents. Before we learned about Maxwell's correction to Ampere's law, the relationship between the magnetic field  $\vec{B}$  and the current i was

$$\oint \vec{B} \cdot d\vec{\ell} = \mu_0 i \quad old \ Ampere's \ law \qquad (29-18)$$

where i was the total electric current flowing through the closed integration loop. Shrinking the integration loop down to infinitesimal size, i.e., going to our differential notation, we get

$$\vec{\nabla} \times \vec{B} = \mu_0 \vec{i} \tag{14}$$

where  $\vec{i}$  is the electric current density. Equation (14), which is missing the  $\partial \vec{E} / \partial t$  term of Maxwell's equation, applies if we can neglect changing electric flux.

In the Physics text, we used the old form of Ampere's law to calculate the magnetic field of a straight wire and of a solenoid. In these examples it was clear that the current i in the wire was the source of the magnetic field.

Let us now compare the equations we have for the magnetic field  $\vec{B}$  (neglecting  $\partial \vec{E} / \partial t$  terms) and for the velocity field  $\vec{v}$  of a constant density fluid. We have

Magnetic field	Velocity Field of Constant Density Fluid	
$\vec{\nabla} \cdot \vec{B} = 0$	$ec{ abla}\cdotec{{ m v}}=0$	
$\vec{\nabla} \times \vec{B} = \mu_0 \vec{i}$	$\vec{\nabla} \times \vec{v} = \vec{\omega}$	(15)

where  $\vec{\omega}$  is the vorticity field of the fluid. If we can interpret  $\mu_0 \vec{i}$  as the source of the magnetic field in the equation  $\vec{\nabla} \times \vec{B} = \mu_0 \vec{i}$ , then by analogy we should be able to interpret the vorticity  $\vec{\omega}$  as the source of the velocity field in the equation  $\vec{\nabla} \times \vec{v} = \vec{\omega}$ .

To be more precise, we will see that the vorticity  $\vec{\omega}$ can be interpreted as the source of any additional velocity beyond the simple potential flow we discussed earlier. If boundary layers, vortices, turbulence, or other derivations from potential flow are present, we can say that vorticity is responsible.

## **Picturing Vorticity**

When we discussed the magnetic field of a current, the current itself was quite easy to picture. It was the flow of electrons along the wire, and for a straight wire this flow of charge produced a circular magnetic field around the wire as shown in Figure (3). We also found from Ampere's law that the strength of the circular magnetic field dropped off as 1/r as we went out from the wire.

In Figure (4) we have drawn a picture of the velocity field of a straight vortex like the one pictured in Figure (23-25) of the Physics text. We observed that the fluid travels in circles around the vortex core. In our funnel vortex we made the core hollow by letting fluid flow out of the funnel, but initially the core contained fluid. We also saw that the fluid flowed faster near the core than far away. The tendency for a fluid vortex is for the velocity field to drop off as 1/r out from the core.

Since the circular velocity field of a straight vortex is similar to the circular magnetic field of a current in a straight wire, we should expect that both fields have similar sources. In Figure (3) the source of the magnetic field is an upward directed current density  $\vec{i}$  in the wire. We therefore expect that the source of the vortex velocity field in Figure (4) should be an upward directed vorticity  $\vec{\omega}$  in the center of the vortex.

Outside the wire, the circular magnetic field drops off as 1/r and has zero curl. If the circular velocity field of the vortex drops off as 1/r outside the core, it must have



A current in a straight wire produces a circular magnetic field around the wire.

zero curl there also. Thus a vortex with a 1/r velocity field outside the core must have all the vorticity  $\vec{\omega}$ concentrated inside the core, just as the current producing the magnetic field is confined to the wire. The vorticity must run up the core as shown in Figure (5). We are beginning to see how the vorticity acts as a source of the velocity field in the same way currents are the source of magnetic fields.





Figure 23-25 Hollow core vortex in a funnel.





#### SOLID BODY ROTATION

Enough of analogies, it is now time to actually calculate the vorticity field  $\vec{\omega} = \vec{\nabla} \times \vec{v}$  of a flow pattern. Our example will be to calculate  $\vec{\omega}$  when  $\vec{v}$  is the velocity field of a solid rotating object.

As an explicit example, imagine that you are looking at the end of a rotating shaft shown in Figure (6). If the shaft has an angular velocity  $\omega_{rot}$ , so that

$$\frac{\mathrm{d}\theta}{\mathrm{d}t} = \omega_{\mathrm{rot}} \tag{16}$$

then at a point p, out at a distance r from the axis of rotation, the velocity is in the  $\hat{\theta}$  direction and given by the formula

$$\vec{v} = \theta r \omega_{rot}$$
 (17)

where the unit vectors  $\hat{\mathbf{r}}$ ,  $\hat{\mathbf{\theta}}$  and  $\hat{\mathbf{z}}$  are for a cylindrical coordinate system are shown in Figure (7).



Figure 6

Figure 7

End of a shaft rotating with an angular velocity  $\omega_{rot}$ .



Unit vectors for a cylindrical coordinate system.

In Chapter 8 of the Calculus text, we wrote down the formula for the curl in cylindrical coordinates. (It can also be found in the Formulary at the end of this text.) Applied to the velocity field  $\vec{v}$ , given by

$$\vec{\mathbf{v}} = \hat{\mathbf{r}}\mathbf{v}_{\mathbf{r}} + \hat{\mathbf{\theta}}\mathbf{v}_{\mathbf{\theta}} + \hat{\mathbf{z}}\mathbf{v}_{\mathbf{z}}$$
(18)

the result is

$$(\vec{\nabla} \times \vec{v})_{\rm r} = \frac{1}{\rm r} \frac{\partial v_z}{\partial \theta} - \frac{\partial v_\theta}{\partial z}$$
 (19a)

$$(\vec{\nabla} \times \vec{v})_{\theta} = \frac{\partial v_r}{\partial z} - \frac{\partial v_z}{\partial r}$$
 (19b)

$$(\vec{\nabla} \times \vec{v})_{z} = \frac{1}{r} \frac{\partial}{\partial r} (rv_{\theta}) - \frac{1}{r} \frac{\partial v_{r}}{\partial \theta}$$
 (19c)

In our example of solid body rotation,  $\vec{v}$  has only a  $\hat{\theta}$  component, and this component  $v_{\theta}(r)$  depends only upon the distance r out from the axis of rotation. Thus  $v_r$ ,  $v_z$ , and  $\partial v_{\theta}/\partial \theta$  and  $\partial v_{\theta}/\partial z$  are all zero and we are left with only the term

$$\left(\vec{\nabla} \times \vec{v}\right)_{z} = \frac{1}{r} \frac{\partial}{\partial r} (r v_{\theta})$$
(20)

You can see that the use of cylindrical coordinates when we have cylindrical symmetry eliminates many terms in the formula for the curl.

#### **Exercise 1**

In the last section, we noted that the circular velocity field of a vortex had zero curl if the velocity drops off as 1/r. This corresponds to a velocity

$$v_{\theta} = \frac{\text{constant}}{r}$$
;  $v_r = v_z = 0$  (21)

Use Equation (19) or (20) to show that  $\vec{\nabla} \times \vec{v} = 0$  for this vortex velocity field.

For solid body rotation, we use  $v_{\theta} = r\omega_{rot}$  to get

$$(\vec{\nabla} \times \vec{v}_{\text{solid body}})_{z} = \frac{1}{r} \frac{\partial}{\partial r} (r v_{\theta})$$
$$= \frac{1}{r} \frac{\partial}{\partial r} (r^{2} \omega_{\text{rot}})$$
$$= \frac{1}{r} (2r \omega_{\text{rot}})$$

$$\left(\vec{\nabla} \times \vec{v}_{\text{solid body}}\right)_{z} = 2\omega_{\text{rot}}$$
<sup>(22)</sup>

Using our notation  $\vec{\nabla} \times \vec{v}_{\text{solid body}} \equiv \vec{\omega}_{\text{solid body}}$ , we get

$$\left(\vec{\omega}_{\text{solid body}}\right)_{z} = 2\omega_{\text{rot}}$$
 (22a)

This is the example we mentioned earlier where the vorticity  $\vec{\omega}$  has a magnitude of exactly twice the rotational velocity  $\omega_{rot}$ .

(It is a challenge to find an intuitive explanation for the factor of 2 difference between the vorticity  $\vec{\omega} = \vec{\nabla} \times \vec{v}$  and the rotational velocity  $\omega_{rot}$ . The analogy is even closer, because when we turned  $\omega_{rot}$  into the vector  $\vec{\omega}_{rot}$  in our discussion of gyroscopes,  $\vec{\omega}_{rot}$  pointed down the rotational axis just as  $\vec{\omega} = \vec{\nabla} \times \vec{v}$  does. I have not met this challenge. After much thought, I have found no satisfactory *intuitive* explanation for the factor of 2. It came in when we differentiated r<sup>2</sup>, but that is not good enough.)



Figure 8

Comparison of the magnetic field of a current in a wire with the velocity field of a fluid core vortex.

The main result from our calculation of the curl for solid body rotation is that the *curl points along the axis of rotation*, and has the *constant magnitude*  $2 \omega_{rot}$  across the entire rotating surface.

#### **Vortex Core**

With our results for the vorticity of solid body rotation, we can see an even closer analogy between the magnetic field of a wire and the vorticity field of a fluid core vortex. The corresponding formulas and field diagrams are shown again in Figure (8).

At the end of Calculus Chapter 8 we studied the magnetic field produced by a uniform current in a wire. We got as the formula for the field inside the wire

$$\vec{B}(r) = \hat{\theta}kr$$
 *inside* (8-66a)

where k was the collection of constants given by

$$k = \frac{\mu_0 t_{\text{total}}}{2\pi R^2} \tag{8-66b}$$

#### Exercise 2

Show that  $\vec{B}$  in Equation (8-66) above obeys the relationship  $\vec{\nabla} \times \vec{B} = \mu_0 i$ .

The magnetic field in Equation (8-66) has the same form as the velocity field for solid body rotation,  $v_{\theta} = r\omega_{rot}$  or

$$\vec{v}_{\text{solid body rotation}} = \hat{\theta}(\omega_{\text{rot}})r$$
 (23)

Thus there will be a complete analogy between the magnetic field of a wire, and a fluid core vortex, if the wire carries a uniform current density i and the vortex core consists of fluid undergoing solid body rotation. In the magnetic field case, the source of the magnetic field is the uniform current in the wire. For the fluid core vortex, the source of the velocity field is the uniform vorticity in the solid body rotating core. Outside the wire and outside the core, both the magnetic field and the velocity field are  $\hat{\theta}$  directed and drop off as 1/r, a field pattern that has zero curl.

#### **STOKES' LAW REVISITED**

For quite a while now we have seen that there are basically two kinds of vector fields. There is what we can call the divergent kind like the electric field of stationary charges that has zero curl. And then there is the rotational kind like the magnetic field and the velocity field of a constant density fluid that has zero divergence. Just as *Gauss' law* played an important role in determining the behavior of divergent fields, we will see that *Stokes' law* has an equally important role in determining the shape and behavior of the rotational kind of vector field. In this section we will take a closer look at Stokes' law, giving it a more physical interpretation than you will find in the mathematics textbooks.

We introduced Stokes' law in Chapter 8 of this text, writing it essentially in the form

$$\oint_{C} \vec{v} \cdot d\vec{\ell} = \int_{S} (\vec{\nabla} \times \vec{v}) \cdot dA \quad Stokes' \ law \qquad (8-14)$$

where  $\vec{v}$  is a vector field, C is some closed contour, and S is the surface bounded by the contour C. We asked you to picture the contour C as being made up of a wire loop, and S the surface of a soap film stretched across the loop. The point was that if you gently blow on a soap film, it can take on various shapes, and Stokes' law applies no matter which shape you consider.



**Figure 8-2 (repeated)** *Example of a surface bounded by a closed path (wire loop).* 

### Total Circulation and Density of Circulation

Because we are going to make extensive use of Stokes' law, we will give special names to the terms in the law. The names are chosen to particularly apply to a velocity field, but can be used in general. First, we will call the line integral of  $\vec{v}$  around a closed path the *total circulation* for the path.

total circulation 
$$\equiv \oint_{C} \vec{v} \cdot d\vec{\ell}$$
 (24)

In addition, we will refer to the vorticity  $\vec{\nabla} \times \vec{v}$  as the *density of circulation* 

density of circulation 
$$\equiv \vec{\nabla} \times \vec{v}$$
 (25)

Then Stokes' law

$$\oint_{\mathbf{C}} \vec{\mathbf{v}} \cdot d\vec{\ell} = \int_{\mathbf{S}} (\vec{\nabla} \times \vec{\mathbf{v}}) \cdot d\mathbf{A}$$

can be stated in words that the *total circulation of the fluid around a closed path C is equal to the density of circulation integrated over any surface bounded by the path.* 

We are using the same terminology one would use in describing a current in a wire. You would say that the total current carried by a wire is equal to the current density integrated over some cross-sectional area of the wire. Why we have introduced this terminology for the velocity field will become clear as we discuss a few examples.

## Velocity Field of a Rotating Shaft, Again

As our first example, let us apply Stokes' law to the velocity field of a rotating shaft, shown in Figure (6) repeated here. Over the area of the end of the shaft we have solid body rotation where the velocity field is  $\hat{\theta}$  directed

$$\vec{v} = \hat{\theta} r \omega_{rot}$$
 (17) repeated

and the vorticity  $\vec{\omega} \equiv \vec{\nabla} \times \vec{v}$  is directed up the axis of the shaft and of magnitude  $2\omega_{rot}$ 

$$\vec{\omega} = \vec{\nabla} \times \vec{v} = \hat{z} 2 \omega_{\text{rot}}$$
 (22) repeated

To apply Stokes' theorem, let the circuit C be the circuit of radius R around the perimeter of the shaft. We then get

$$\oint_{C} \vec{v} \cdot d\vec{\ell} = \oint v_{\theta}(d\ell)_{\theta}$$
(26)

At the perimeter,  $v_{\theta} = R\omega_{rot}$ , and  $(d\ell)_{\theta} = Rd\theta$ , to give

$$\oint \vec{v} \cdot d\vec{\ell} = \int_{0}^{2\pi} (R\omega_{rot})(Rd\theta)$$
$$= R^{2}\omega_{rot} \int_{0}^{2\pi} d\theta = 2\pi R^{2}\omega_{rot}$$

Thus the total circulation of the shaft is given by



**Figure 6 (repeated)** End of a shaft rotating with an angular velocity  $\omega_{rot}$ .

Stokes' theorem states that this total circulation should be equal to the density of circulation  $\vec{\nabla} \times \vec{v}$ integrated over the area of the shaft. We know that for solid body rotation

$$\frac{\text{density of}}{\text{circulation}} = \vec{\nabla} \times \vec{v} = \vec{\omega} = \hat{z} \, 2 \, \omega_{\text{rot}}$$
(28)

This density, of magnitude  $\omega_z = 2 \omega_{rot}$ , is constant over the area of the shaft, thus the integral of the density is simply

$$\int_{S} (\vec{\nabla} \times \vec{v}) \cdot d\vec{A} = \int_{S} \omega_{z} dA_{z}$$
$$= \omega_{z} \int_{S} dA_{z} = \omega_{z} \pi R^{2}$$
$$= \pi R^{2} (2\omega_{rot})$$
(29)

Comparing Equations (27) and (29), we see that the total circulation is, as expected, equal to the density integrated over the area of the shaft.

### Wheel on Fixed Axle

Before you think everything is too obvious, let us consider a more challenging example. Suppose we have a wheel of radius R, rotating on a fixed axle of radius  $R_{axle}$ , as shown in Figure (9). The velocity field for this example is

$$\vec{v} = 0$$
  $r < R_{axle}$   
 $\vec{v} = \hat{\theta} r \omega_{rot}$   $R_{axle} < r < R$  (30)



**Figure 9** *Wheel rotating on a stationary axle.* 

To apply Stokes' law again, let C be a circuit of radius R about the perimeter of the wheel. The total circulation is the same as before, namely

total  
circulation = 
$$\int_{C} \vec{v} \cdot d\vec{\ell} = (R\omega_{rot})(2\pi R)$$
  
=  $\pi R^2 (2\omega_{rot})$  (31)

When we measure the total circulation around the wheel, the result is uniquely determined by the value of  $\vec{v}$  out at the circuit C. It makes no difference whatever whether the axle inside is turning or not.

But when we integrate the density of circulation  $\vec{\nabla} \times \vec{v}$  over the area of the wheel, we have a problem. Over the wheel  $\vec{\nabla} \times \vec{v} = \hat{z} \, 2 \, \omega_{rot}$  as before, but  $\vec{\nabla} \times \vec{v} = 0$  over the axle. It appears that we have lost an amount of circulation  $(2 \, \omega_{rot})(\pi R_{axle}^2)$ , and that Stokes' law fails.

Mathematics textbooks would say that we did not apply Stokes' law correctly. You will find statements like "Stokes' law applies only to singly connected surfaces" or "you have to add a cut". **Don't believe it!** Stokes' law applies quite generally, and you do not need so called cuts. What went wrong in this example is not Stokes' law, it is that we did not look carefully enough.

Suppose Figure (9) represented the wheel on a railroad car. Look carefully at the boundary between the wheel and the axle and what do you find? *Roller bearings!* As the wheel rotates on the axle, the roller bearings really spin. The circulation that we lost in the axle is now located in the roller bearings, and in the velocity field of the oil lubricating the bearings.



**Figure 9a** Wheel with roller bearings rotating on a stationary axle.

You might be a bit worried about this explanation. After all, a fixed amount of circulation, namely  $(2\omega_{rot})(\pi R_{axle}^2)$  was lost when we stopped the axle from rotating. But the space where the roller bearings reside, between the axle and the wheel can be made as thin as we want, reducing the area of the bearings that we integrate  $\vec{\nabla} \times \vec{v}$  over. If we make the area of the bearings go to zero, can we still get a finite amount of circulation  $(2\omega_{rot})(\pi R_{axle}^2)$  when we integrate over this vanishing area?

The answer is yes. Look what happens to roller bearings as we make the diameter of the bearings smaller and smaller. They have to spin faster and faster so that they roll smoothly between the axle and the wheel. As we decrease the thickness of the bearings, we increase the vorticity  $\vec{\nabla} \times \vec{v}$  in the bearings in just such a way that the integral of  $\vec{\nabla} \times \vec{v}$  over the bearings remains constant. In the mathematical limit that the thickness of the bearings goes to zero, we end up with a delta function of vorticity spread around the perimeter of the axle. This delta function of vorticity is called a *vortex sheet*. When you correctly account for vortex sheets, you can always make sense of Stokes' law without caveats relating to singly connected surfaces or cuts.

## A Conservation Law for Vorticity

Imagine that our solid shaft of Figure (6) represented a wheel and axle where the axle was rotating with the wheel. Then the axle would have vorticity of magnitude  $2\omega_{rot}$  just like the wheel. Now suppose we grab hold of the axle to stop it from rotating, giving us the velocity field shown in Figure (9). By stopping the axle from rotating, we did not destroy the vorticity, we just moved it out to the roller bearings or vortex sheet. For a given total circulation around the rim of the wheel, we cannot create or destroy vorticity within, only move it around. With a given total circulation, we have a conserved amount of vorticity within. In this sense, Stokes' law provides us with a conservation law for vorticity. (In Appendix 2 of Chapter 13, we show you a more general, three dimensional law for the conservation of vorticity.)

## **CIRCULATION OF A VORTEX**

In an ideal straight vortex like the one we pictured in Figure (8) more or less redrawn here as Figure (10), the vorticity is concentrated in the core and we have a curl free 1/r velocity field outside the core. It is traditional to use the Greek letter  $\kappa$  (kappa) to designate the total circulation of the vortex.

$$\oint_{\text{over any area that includes the vortex core}} \vec{v} \cdot d\vec{\ell} = \kappa \quad total circulation or strength of a vortex (32)$$

Evaluating the integral around a circle outside the core gives

$$\oint \vec{v} \cdot d\vec{\ell} = 2\pi r v_{\theta} = \kappa$$

$$\boxed{v_{\theta} = \frac{\kappa}{2\pi r}} \qquad \begin{array}{c} \text{velocity field} \\ \text{of } a \\ \text{straight vortex} \end{array} (33)$$

This is the formula for the velocity field of a straight vortex, outside the core. For shorthand, we sometimes use  $\bar{\kappa} = \kappa/2\pi$  just as we used  $\hbar = h/2\pi$  in quantum mechanics, giving

$$\begin{bmatrix} v_{\theta} = \frac{\bar{\kappa}}{r} \\ r \end{bmatrix} \qquad \begin{array}{c} velocity \ field \\ of \ a \\ straight \ vortex \end{array} \tag{33a}$$

Note that talking about the total circulation  $\kappa$  of a vortex, we know that when there is cylindrical symmetry, the velocity field  $v_{\theta}$  outside the core is  $\kappa/r$  independent of the structure of the core. The core can be a fluid core with solid body rotating fluid inside, or be a hollow core vortex like the funnel vortex of Figure (23-25). With a solid body rotating core the vorticity  $\vec{\omega}$  is spread uniformly across the core. With a hollow core vortex, we can think of the vorticity as being in a vortex sheet around the core.

We have a similar situation for the magnetic field of a straight wire. In a normal wire, there is a more or less uniform current density in the wire which produces a magnetic field of strength B<sub> $\theta$ </sub> =  $\mu_0 I_{total}/2\pi r$ outside. In some superconducting wires, those made from the so called type 1 superconductors like lead and tin, the electric current flows very near the surface of the wire with no current farther inside. This surface sheet of current still produces the same magnetic field B<sub> $\theta$ </sub> =  $\mu_0 I_{total}/2\pi r$  outside.



#### Figure 10

The total circulation  $\vec{\kappa}$  of the vortex is related to the velocity field  $\vec{\nu}$  the same way the total current  $\vec{i}_{tot}$  is related to the magnetic field  $\vec{B}$ . (For straight vortices, we often think of  $\vec{\kappa}$  as a vector pointing in the direction of  $\vec{\omega}$ , as shown above.)

## **QUANTUM VORTICES**

We are now ready to deal with the failure of Landau's prediction that superfluid helium could only undergo potential flow, with the consequence that helium in a bucket could not rotate. The appearance of a parabolic surface on a rotating bucket of superfluid helium is experimental evidence that vorticity is present in the fluid despite Landau's prediction.

Feynman solved the problem by proposing that most of the fluid in a rotating bucket of superfluid helium was in fact undergoing potential flow, and that all the vorticity that was responsible for the curved surface was contained in little quantized vortices.

As we have mentioned in the Physics text, a single quantized vortex can be pictured as a giant Bohr atom where all the superfluid atoms taking part in the vortex flow have one unit of angular momentum  $\hbar$  about the vortex core. The angular momentum of an atom out at a distance r from the core, moving at a speed v<sub> $\theta$ </sub>, is

$$L \frac{\text{angular}}{\text{momentum}} = m_{\text{He}} v_{\theta} r$$
(34)

where  $m_{He}$  is the mass of a helium atom. If we set the angular momentum L equal to Planck's constant  $\hbar$ , and solve for  $v_{\theta}$ , we get

$$L = \hbar = m_{He} v_{\theta} r$$

$$v_{\theta} = \frac{\hbar}{m_{He} r}$$
(35)

We immediately see that the velocity field outside the core drops off as 1/r which is potential flow.



#### Figure 11

Each atom in a quantum vortex has one unit of angular momentum about the vortex core.

The 1/r velocity field cannot continue in to r = 0; there has to be a core that is not potential flow. There are two questions that need to be settled by experiment. One is how big is the core radius  $r_{core}$ , and the second is whether the core is hollow, or filled with rotating fluid. The answer to the first question is rather amazing. Under most circumstances the core is about as small as it can get, *about one atomic diameter*. That makes it difficult to answer the second question; it is hard to tell what is inside a tube only one atomic diameter across.

## **Circulation of a Quantum Vortex**

One thing we can do immediately from Equation (35) is to calculate the total circulation  $\kappa$  of a quantum vortex. Remembering that  $\hbar = h/2\pi$  we have

$$v_{\theta} = \frac{h}{m_{He}r} = \frac{h}{(2\pi r)m_{He}}$$
$$(2\pi r)v_{\theta} = \frac{h}{m_{He}}$$

But  $2\pi r v_{\theta}$  is simply the integral of  $v_{\theta}$  around a circle centered on the core. Thus we have

$$\oint \vec{\mathbf{v}} \cdot d\vec{\ell} = 2\pi r \mathbf{v}_{\theta} = \kappa$$

$$\boxed{\kappa = \frac{h}{m_{\text{He}}}} \qquad \begin{array}{c} \text{circulation of } a \\ \text{quantum vortex in} \\ \text{superfluid helium} \end{array} (36)$$

#### **Rotating Bucket of Superfluid Helium**

If you have a rotating bucket of normal fluid, the fluid will end up rotating with solid body rotation with constant vorticity  $\vec{\omega} = \hat{z} 2\omega_{rot}$ . The total circulation  $\kappa_{total}$  of all the fluid in the bucket will be

total circulation of  
fluid in rotating  
bucket = 
$$\int_{\substack{\text{bucket}\\\text{surface}}} (\vec{\nabla} \times \vec{v}) \cdot dA$$

 $\kappa_{\text{total}} = (2\omega_{\text{rot}})(\pi R^2_{\text{bucket}})$ (37)

For solid body rotation, this vorticity is spread uniformly across the bucket.

Feynman proposed that a rotating bucket of superfluid helium would have the same total circulation  $\kappa_{total}$ , but that the vorticity, instead of being spread throughout the fluid, would be contained in a bundle of quantized vortex cores. This difference between the classical and quantum picture is indicated in Figure (12).

Because the core of a quantum vortex is so small, and because all the fluid between the cores is undergoing potential flow, you can see that Landau was almost right. But the quantum cores allow vorticity to be spread throughout the bucket, roughly imitating solid body rotation, and give rise to a nearly parabolic surface.



#### Figure 12

Comparison of solid body rotation with a bundle of quantized vortices. (We have not tried to reproduce the exact shape of the surface when vortices are present.) Between the vortices the flow is potential, but the rough shape of the surface is parabolic. We can easily calculate the number of quantized vortices required to imitate solid body rotation. From Equation (37), we saw that the total circulation of the bucket was  $\kappa_{total} = (2\omega_{rot})(\pi R^2_{bucket})$ . Each quantum vortex supplies a circulation h/m<sub>He</sub>. If we have N quantum vortices, their total circulation will be Nh/m<sub>He</sub>. Equating these two numbers gives

$$\kappa_{\text{total}} = (2\omega_{\text{rot}})\pi R^2_{\text{bucket}} = N \frac{h}{m_{\text{He}}}$$

Solving for N, and then dividing by the area of the bucket, gives us the number n of quantized vortices per unit area.

$$n = \frac{N}{\pi R_{bucket}^2} = \frac{2\omega_{rot}m_{He}}{h}$$
(38)

To see what the density is of quantized vortices needed to imitate solid body rotation, let us use CGS units where the unit area is  $1 \text{ cm}^2$ , and solve for an angular velocity  $\omega_{\text{rot}}$  of one radian/second which is about 1/6 of a revolution per second. We have

$$\begin{split} &\omega_{rot}~=~1\\ &m_{He}~=~4\times1.67\times10^{-~24}gm~~\left(\textit{4 proton masses}\right)\\ &h~=~6.62\times10^{-~27} \end{split}$$

We get for the vortex density n

$$n = \frac{2\omega_{rot}m_{He}}{h}$$
$$= \frac{2 \times 4 \times 1.67 \times 10^{-24}}{6.62 \times 10^{-27}}$$
$$= 2020 \text{ lines/cm}^2$$

If these lines were in a rectangular array, there would be  $\sqrt{n}$  lines on each side of a square centimeter

 $\sqrt{n} = 45 \text{ lines/cm}$ 

The spacing between lines would be  $1\sqrt{n}$ 

$$1\sqrt{n} = .022 \text{ cm/line}$$
  
= .22 millimeters/line (39)

Thus to imitate solid body rotation with an array of quantized vortices in superfluid helium, the quantum vortices have to be .22 millimeters apart when the rotational velocity is 1 radian per second.

For a number of years after Feynman's explanation of the curved surface on a bucket of superfluid helium, there was a considerable effort to see if quantum vortices really exist in the superfluid. The most conclusive evidence for their existence, with the predicted circulation  $\kappa = h/m_{He}$ , came from experiments by Rayfield and Reif using charged vortex rings. A few years later Richard Packard at Berkeley succeeded in actually photographing the vortices in a rotating bucket of helium. He did this by loading up the vortex lines with electrons, and then firing the electrons into a film placed at the surface of the liquid. The result is shown in Figure (13) for various rotational speeds.

What Feynman and others have shown is that the flow pattern with quantized vortices is a wave pattern for the helium atoms in the bucket. It is the lowest energy solution of a wave equation, subject to the boundary condition that the atoms near the surface of the bucket are moving with a velocity nearly equal to the velocity of the bucket. Although we have used the terminology of classical fluid dynamics, we are describing a quantum mechanical phenomenon. What is remarkable is that we are seeing quantum mechanical phenomena on a large human scale, not just an atomic scale. You can see a separation of .22 millimeters without the use of a microscope.

#### Exercise 3 - A Superfluid Gyroscope

Counting vortices in a bucket of superfluid helium can be a sensitive way of detecting rotation. Suppose a bucket of helium were placed at the North Pole. How many vortices per cm<sup>2</sup> would there be in the bucket due to the rotation of the earth?





Packard's photograph of vortex lines in rotating superfluid helium. As the rotational speed is increased, more quantum vortices appear. Angular velocities range up to half a radian per second. (The camera was rotated with the helium and many exposures were taken to build up the image. The slight jiggling of the vortices between exposures spread the vortex images out a bit.)

### **Bose-Einstein Condensates**

Since 1995, it has been possible to create a new kind of superfluid, consisting of a small drop of gas cooled to temperatures in the range of a millionth of a kelvin. What happens to the gas atoms at these temperatures is that they can come together and "condense" into a single quantum mechanical wave pattern. The process is not unlike photons condensing into a single wave pattern in a laser beam. For the gas atoms the result is a liquid-like drop with superfluid properties.

It is called Bose-Einstein condensation because back in the 1920's, Einstein predicted this effect, basing his ideas on the work of the Indian physicist Nath Bose. It turns out that atoms or objects that have integer spin like to congregate into a single quantum wave pattern if the temperature is low enough, i.e., if the pattern is not disturbed by thermal effects. Examples of integer or zero spin objects that do this are photons that form laser beams, Helium 4 atoms that form superfluid helium, and electron pairs that become a superconductor.

In 1999, a group at the École Normale Supérieure in Paris succeeded in rotating a drop of rubidium atoms and photographing the quantized vortices as they appeared. Due to the weak attraction between the rubidium atoms, the vortex cores are some 5000 times bigger than the core of a superfluid helium vortex, but have the same circulation h/matom. Photographs of the drop, with 0, 1, 8, and 13 vortices are seen in Figure (14). Figure (15) is a computer simulation of the vortex core structure of a drop with four vortices passing through the drop, and two forming at the edge.







Figure 15

## THE VORTICITY FIELD

So far we have described vorticity as something we look for in a vortex core or something that characterizes solid body rotation. In this section we will treat the vorticity  $\vec{\omega} = \vec{\nabla} \times \vec{v}$  as a dynamic field that has field lines and can behave much like the other vector fields we have been discussing.

The singular property of vorticity is that it always has identically zero divergence

$$\vec{\nabla} \cdot \vec{\omega} = \vec{\nabla} \cdot (\vec{\nabla} \times \vec{v}) \equiv 0 \tag{40}$$

because the divergence of a curl is identically zero. (See the vector identities.) This means that vorticity is always a solenoidal field without sources or sinks.

We defined a field line of the velocity field as a small flow tube, like those seen in Figure (23-3) reproduced below. Similarly, we define a vortex line as a small flow tube of vorticity. The total flux of vorticity in the flow tube is by definition, the circulation  $\kappa$  of that tube. As a reminder, this comes from Stokes' law

flux of 
$$\vec{\omega}$$
  
in a vortex =  $\int_{\substack{\text{surface}\\ \text{across}\\ \text{tube}}} \vec{\omega} \cdot dA = \int_{S} (\vec{\nabla} \times \vec{v}) \cdot dA$   
=  $\oint_{\substack{\text{around}\\ \text{around}}} \vec{v} \cdot d\ell = \kappa_{\text{tube}}$  (41)

Because the vorticity  $\vec{\omega}$  is solenoidal, the flux tubes or lines of  $\vec{\omega}$  cannot start or stop inside the fluid. Vortex lines can only start or stop on the fluid boundaries, or close on themselves within the fluid. Two examples are the straight vortices we have been discussing which run from the bottom of a container to the top, and a vortex ring where the vortex lines go around and close on themselves like the magnetic field lines around a wire. A smoke ring is the classic example of a vortex ring.

Figure 23-3 Flow tubes bounded by streamlines. We define a field line as a small flow tube.



## **HELMHOLTZ THEOREM**

In 1858 Heinrich Helmholtz discovered a remarkable theorem related to vortex motion. He discovered that *when all the forces acting on fluid particles are conservative forces, i.e., force fields that have zero curl, vortex lines move with the fluid particles*. Gravity is an example of a conservative force, viscous forces are not. If viscosity can be neglected and only gravity is acting on the fluid, vortex lines and fluid particles move together.

To emphasize this point, in the absence of non conservative forces, we can say that the fluid particles become trapped on vortex lines, or we can say that vortex lines become stuck on and have to move with the fluid particles. *To move vorticity onto or off a fluid particle requires a non conservative force like viscosity*.

## The Two Dimensional "Vortex Ring"

The simplest illustration of Helmholtz's theorem is the behavior of a vortex ring where the vortex lines go around a circle and close on themselves. The most well known example of a vortex ring is the smoke ring.

Before we discuss circular vortex rings, we will consider the simpler example of two oppositely oriented straight vortices which form what is often called a *two dimensional (2D) vortex ring*. A view down upon the two vortices, showing their independent velocity fields, is shown in Figure (16). The total velocity field of these two vortices is the vector sum of the fields from each vortex. Notice that the upper vortex has a forward velocity field at the lower vortex core. If Helmholtz's theorem is obeyed, then this upper velocity field must be moving the vortex lines in the lower core forward. Likewise the velocity field of the lower vortex must move the core of the upper vortex forward. As a result this two dimensional vortex configuration is a self propelled, forward moving object.

We can easily calculate the forward speed of our 2D vortex ring. The velocity field of a vortex of circulation  $\kappa$  was given by Equation (33a) as

$$v_{\theta} = \frac{\kappa}{r}$$
;  $\kappa = \frac{\kappa}{2\pi}$  (33) repeated

If the separation of the vortices is d, then the speed of the fluid at the opposite core, and therefore the speed of the ring will be



You can see that the ring moves faster (1) if the circulation  $\bar{\kappa}$  is increased, or (2) if the vortices are closer together.



Velocity fields of two oppositely oriented straight vortices.

#### **The Circular Vortex Ring**

For a circular, or 3D vortex ring, the vortex core has the shape of a doughnut. If we look at the velocity field in a plane that slices through the doughnut, as shown in Figure (17), the result is in many ways similar to the velocity field of the 2D vortex in Figure (16). In particular the velocity field of the top part of the ring moves the bottom part of the ring forward, while the field of the bottom of the ring moves the top forward.

In addition, the smaller the ring, the faster it moves. If the ring has a circulation  $\kappa$  and diameter d, the speed of the ring is approximately given by the same equation  $v_{ring} = \bar{\kappa}/d$  that applied to the 2D vortex.

The actual velocity field of a vortex ring has the same shape as the magnetic field of a circular current loop, (provided the current density in the wire has the same shape as the vorticity in the vortex core). It is a classic and rather nasty problem to calculate the precise shape of this field. When we get a more accurate answer for the speed of the ring, we end up with additional terms, one of which involves the logarithm of the core radius. This logarithm would go to infinity if we tried to make the core radius zero, but the term becomes small for reasonable core radii. We do not need to worry about these small additional terms now. The analogy to the behavior of the two dimensional ring is good enough.



**Figure 17** *Velocity field in a slice through a vortex ring.* 

## **Smoke Rings**

In several ways the smoke ring provides a superb illustration of Helmholtz's theorem. In the days when smoking was popular and thought to be harmless, it was a common stunt to blow a smoke ring. Today we would rather create smoke rings using the apparatus shown in Figure (18). The apparatus is simple, and the rings are better.

Start with a cardboard box, cut a fairly large hole in the front as shown, and replace the back side with a rubber sheet. Fill the box with smoke, and hit the rubber sheet with your hand. A beautiful ring will emerge, like the one shown in Figure (19).

(If titanium tetrachloride solution available, you can get a denser smoke ring by squirting this liquid around the perimeter of the hole in the box. The titanium tetrachloride quickly turns to titanium dioxide smoke and hydrochloric acid. The titanium dioxide is a coloring agent for white paint, and the hydrochloric acid is obnoxious to deal with, but the resulting rings are quite good.)



**Figure 18 a,b** Front and back of apparatus for creating smoke rings.



**Figure 18 c** Smoke at hole due to titanium tetrachloride.

The most impressive feature of the smoke rings created by our box is how stable they are. They move in a straight line, at constant speed, without changing their shape, just as predicted by our analysis of the two and three dimensional vortex rings. If you hit the rubber sheet harder, you add more circulation  $\kappa$  to the rings, and they travel faster. You can experiment with different size holes in the box, seeing that smaller rings travel faster than larger ones.

One of the interesting predictions that you can think about and try to observe is the following. If a faster ring approaches a slower one in front of it, the velocity field of the front ring will tend to make the back ring smaller and thus move still faster. Conversely, the velocity fields of the back ring should expand the front ring making it move more slowly. (Sketch the velocity fields yourself to check this prediction.) As a result, if the back ring is aimed right at the front one, the smaller back ring should shoot through the larger front ring, becoming itself the front ring. If the rings have not bumped into each other, tangled and destroyed themselves (the usual case), then the new back ring will be squeezed in size, the front ring expanded, and the process repeated. This is a famous prediction, but I have not seen it carried out very well.

While the motion of a smoke ring represents a successful prediction of Helmholtz's theorem, the fact that the smoke ring is so sharply defined, escap-

ing from the amorphous cloud of smoke around the cardboard box, is an even more dramatic prediction of the theorem. When we hit the back of the box to create the ring, air was expelled out through the hole in the front. The vortex ring was created at the perimeter of the hole from air that contained smoke particles. *These smoke particles in the vortex core become attached to the vortex lines in the core* and have to move with the core. As the vortex ring moves out of the box, it carries the trapped smoke particles in its core and leaves the rest of the smoke behind.

## **Creating the Smoke Ring**

The reason why is as follows. Before we hit the rubber sheet at the back of the box, all the air in the box was at rest and contained no vorticity. If Helmholtz's theorem strictly applied, then a vortex line could not move onto fluid particle that initially had no vorticity.

As we mentioned earlier, Helmholtz's theorem applied *if only conservative forces* (like gravity) were acting on the fluid. But gravity is not the only force acting on the particles of air in our smoke ring apparatus. Air is a slightly viscous fluid, and viscous forces in a fluid are not curl free conservative forces. Viscous forces move a vortex line onto fluid particle and create a vortex core.



**Figure 19** *Two smoke rings after they have collided.* 

## Calculus 2000-Chapter 13 Introduction to Fluid Dynamics

One should think of this chapter as an introduction to fluid dynamics. In it we derive the basic equations for the behavior of the velocity field  $\vec{v}$  and the vorticity field  $\vec{\omega}$  in a constant density fluid. We begin by applying Newton's second law to a fluid particle to obtain what is known as the Navier-Stokes equation. This equation for the velocity field  $\vec{v}$  serves as the fundamental equation of fluid dynamics.

Taking the curl of the Navier-Stokes equation gives us the basic equation for the dynamics of the vorticity field  $\vec{\omega}$ . From that equation we derive the Helmholtz theorem, and an extension of the Helmholtz theorem that deals with the effect of non potential forces acting on fluid cores. The extended Helmholtz theorem is used in the analysis of the experiments of Rayfield and Reif who first measured the circulation  $\kappa$  and core radius (a) of a quantized vortex in superfluid helium. We end the regular part of the chapter with a discussion of the **Magnus effect** and the pseudo force called the **Magnus force** that appears in all the vortex dynamics literature. There are two major appendices to this chapter. Appendix 1 deals with the use of component notation in vector equations. This includes the Einstein summation convention, and emphasizes the use of the permutation tensor  $\varepsilon_{ijk}$  for calculating vector cross products. There we show you an easy way to derive vector identities involving cross products.

The second appendix shows how you can interpret the dynamical behavior of the vorticity field as a conserved two dimensional flow of vorticity. Appendix 2 begins with an intuitive derivation of that result, a derivation that requires little mathematical background. (It can be explained at dinner parties.) However deriving the formula for the conserved vortex current requires the use of the permutation tensor  $\varepsilon_{ijk}$ , which is why we delayed this discussion until after Appendix 1.

The use of vortex currents turns out to be a particularly effective way to handle vortex motion. We use it, for example, to derive the Magnus force equation for curved fluid core vortices, a result that has not been obtained any other way.

## THE NAVIER-STOKES EQUATION

When we apply Newton's second law  $\vec{F} = d\vec{p}/dt$  to a particle like a baseball, the analysis is fairly simple. With  $\vec{p} = m\vec{v}$  for the baseball, if m is constant, the result is  $\vec{F} = md\vec{v}/dt$ . In particular, if  $\vec{v} = \text{constant}$ , then  $d\vec{v}/dt = 0$  and  $\vec{F} = 0$ .

Applying Newton's second law to a fluid is more complicated. Even if we have a steady flow where  $\vec{v} = \text{constant}$ , the fluid particles themselves will be accelerating when the streamlines go around a corner or the flow tubes become narrower or wider. Some net force acting on the fluid particles is required to produce this acceleration. If the flow is not steady, if  $\partial \vec{v} / \partial t$  is not zero, an additional force is required to produce this change in the velocity field. The first problem you encounter in the study of fluid mechanics is to correctly evaluate the acceleration of the fluid particles taking both of these effects into account.

What we will do is to consider a volume V of fluid bounded by a closed surface S'. The surface S' is special in that it moves with the fluid particles. As a result the same fluid particles remain inside V as the fluid moves about. We will then calculate the rate of change of the total momentum of these fluid particles and equate that to the total force acting on the particles within V. Following this procedure we will end up with a differential equation called the *Navier-Stokes* equation which is very successful in describing the behavior of fluids.

(In most textbooks you will find what looks to be a simpler derivation of the Navier-Stokes equation. Our derivation involves volume and surface integrals, while the textbooks make what looks like simpler arguments using what is called a **substantive derivative**. When the textbook arguments are applied to non constant density fluids, you also find some talk about what should be included inside the substantive derivative and what should not. It almost seems that one includes only those terms that give the right answer.

By using surface and volume integrals, our focus remains on the application of Newton's second law to the fluid particles with no ambiguities of interpretation.)

#### **Rate of Change of Momentum**

As we mentioned, we will consider a volume V of fluid whose surface S' moves with the fluid particles. As a result the same particles remain inside the volume V. We then equate the rate of change of the total momentum of these particles to the total force acting on them. The main problem involves calculating the rate of change of the momentum of the particles in a volume whose surface is moving.

Suppose we have a volume V(t) that is now, at time (t), bounded by a surface S'(t) ( shown in Figure 1). If the fluid has a density  $\rho$  and the velocity field of the fluid is  $\vec{v}$  then the total momentum  $\vec{P}_V(t)$  of the fluid in V(t) is

$$\vec{P}_{V}(t) = \int_{V(t)} \vec{p}(\vec{x},t) d^{3}V$$
 ;  $\vec{p} = \rho \vec{v}$  (1)

At this point we are even allowing the density to vary, so that both  $\rho$  and  $\vec{v}$  can be functions of space and time.

A short time  $\delta t$  later, the surface will have moved to  $S'(t+\delta t)$  and the volume becomes  $V(t+\delta t)$  as shown in Figure (2).

At this later time, the momentum of the fluid particles will be



**Figure 1** *The volume V bounded by the surface S' at time (t).* 



**Figure 2** *The volume V a short time St later.* 

$$\vec{P}_{V}(t+\delta t) = \int_{V(t+\delta t)} \vec{p}(t+\delta t) d^{3}V \qquad (2)$$

The change  $\delta \vec{P}_V$  in momentum of the fluid particles as time goes from (t) to  $(t+\delta t)$  is

$$\begin{split} \delta \vec{P}_{V} &= \vec{P}_{V}(t + \delta t) - \vec{P}_{V}(t) \\ &= \int_{V(t + \delta t)} \vec{p}(t + \delta t) d^{3}V - \int_{V(t)} \vec{p}(t) d^{3}V \end{split} \tag{3}$$

We can do a Taylor series expansion of  $\vec{p}(t+\delta t)$  to get

$$\vec{p}(t+\delta t) = \vec{p}(t) + \frac{\partial \vec{p}}{\partial t} \delta t + O(\delta t^2)$$
(4)

This gives

$$\begin{split} \delta \vec{P}_{V} &= \left[ \int_{V(t+\delta t)} \vec{p}(t) d^{3}V - \int_{V(t)} \vec{p}(t) d^{3}V \right] \\ &+ \delta t \int_{V(t+\delta t)} \frac{\partial \vec{p}}{\partial t} d^{3}V + 0(\delta t^{2}) \end{split} \tag{5}$$

From Figure (2), we see that much of the same volume is included in both  $V(t+\delta t)$  and V(t). Thus, in the square brackets in Equation (5), the integral of  $\vec{p}(t)$  over the common volume cancels, and what we want is an integral of  $\vec{p}(t)$  over the volume that the fluid has entered during the time  $\delta t$ , minus the integral of  $\vec{p}(t)$  over the volume the fluid has left during  $\delta t$ .

In Figure (3a) we show part of the region between S'(t) and  $S'(t+\delta t)$  where the fluid has entered during  $\delta t$ . Consider a particle at point (1) at time t, moving at a velocity  $\vec{v}_1$ . In the short time  $\delta t$  it moves a distance  $\vec{v}_1 \delta t$  as shown.

Now let  $d\vec{A}_1$  be an element of the surface S'(t) at point (1). The standard convention is that a surface element  $d\overline{A}$  points perpendicularly *out* of a closed surface. Thus  $dA_1$  points out of surface S'(t) as shown.

A time  $\delta t$  later, the surface element  $d\vec{A}_1$  will have moved out to the surface S'(t + $\delta$ t), sweeping out a volume  $\delta V_1$  given by

$$\delta \mathbf{V}_1 = (\vec{\mathbf{v}}_1 \delta \mathbf{t}) \cdot \mathbf{d} \vec{\mathbf{A}}_1 \tag{6}$$

You can see that the dot product is appropriate, for if  $\vec{v}_1$  and  $d\vec{A}_1$  are parallel, we have a right circular cylinder of volume  $(v_1 \delta t dA_1)$ . The volume is zero if  $\vec{v}_1$  and  $d\vec{A}_1$  are perpendicular, and negative if oppositely oriented.

In Figure (3b) we show part of the region between S'(t) and S'(t + $\delta$ t) where the fluid in S'(t) has left during the time  $\delta t$ . The diagram is the same as Figure (3a) except that the vector  $d\vec{A}_2$  pointing out of S'(t) is pointing essentially opposite to the vector  $\vec{v}_2$ . In the formula  $\delta V_2 = (\vec{v}_2 \delta t) dA_2$ , the dot product  $\vec{v}_2 \cdot d\vec{A}_2$  and therefore  $\delta V_2$  is negative in the region where the fluid is leaving.

As a result, if we calculate the integral of  $\vec{p}(t)\delta V$ over both the volumes in Figures (3a) and (3b), we get an integral of  $\vec{p}(t)$  over the region the fluid is entering, minus the integral of  $\vec{p}(t)$  over the region the fluid is leaving. This just gives us the quantity in the square brackets in Equation (5)



The volume element  $\delta V_1 = \vec{v} \delta t \cdot d\vec{A}_1$ into which the fluid is flowing.

Figure 3a



#### **Figure 3b** The volume element $\delta V_2 = \vec{v} \delta t \cdot d\vec{A}_2$ out of which the fluid is flowing.

We get

$$\begin{bmatrix} \int \vec{p}(t) d^{3}V - \int \vec{p}(t) d^{3}V \\ V(t+\delta t) & V(t) \end{bmatrix}$$
  
= 
$$\int \int \vec{p}(t) (\delta V) = \int \vec{p}(t) (\delta t \vec{v} \cdot d\vec{A}) \qquad (6)$$
  
and leaving regions

By integrating over the entire area S'(t) we have included both the entering and leaving regions.

Using Equation (6) for the square brackets in Equation (5) gives

$$\delta \vec{P}_{V} = \delta t \int_{S'(t)} \vec{p}(t)(\vec{v} \cdot d\vec{A}) + \delta t \int_{V(t+\delta t)} \frac{\partial \vec{p}(t)}{\partial t} d^{3}V$$
(7)

plus terms of the order  $\delta t^2$ . At this point, we have everything expressed at the time (t) except the volume of integration in the  $\partial \vec{p}/\partial t$  term. If we integrated over the volume V(t) instead of V(t +  $\delta t$ ), we would be incorrectly handling the integral of  $\partial \vec{p}/\partial t$  over the narrow difference volume of thickness  $\vec{v}\delta t$ . Since the  $\partial \vec{p}/\partial t$  term already has a factor  $\delta t$ , this would lead to an error of order  $\delta t^2$  which we can ignore.

Replacing  $V(t + \delta t)$  by V(t) in the volume integral, and dividing through by  $\delta t$  gives

$$\frac{\delta \vec{P}_{V}}{\delta t} = \int_{V(t)} \frac{\partial \vec{p}}{\partial t} d^{3}V + \int_{S'(t)} \vec{p}(t)(\vec{v} \cdot d\vec{A})$$
(8)

We now have all quantities in our formula for  $\delta \vec{P}_V / \delta t$  expressed at the time (t).

We have one more step before we are finished with the  $\delta \vec{P}_V / \delta t$  term. We want to convert the surface integral to a volume integral.

We have already had some experience converting surface to volume integrals back in Chapter 7 on divergence. There we derived the *divergence theorem* 

$$\int_{S} \vec{E} \cdot d\vec{A} = \int_{V} \vec{\nabla} \cdot \vec{E} d^{3}V$$
(7-21)

where  $\vec{E}$  is any vector field, and the surface S bounds the volume V.

In Equation (8), we have something that looks more complex than the surface integral in (7-21), because of the presence of the extra vector  $\vec{p}$ . To handle this let us define three fields  $\vec{E}_1$ ,  $\vec{E}_2$  and  $\vec{E}_3$  by

$$\vec{E}_1 = p_x \vec{v}$$
;  $\vec{E}_2 = p_y \vec{v}$ ;  $\vec{E}_3 = p_z \vec{v}$  (9)

Then we get

$$\int_{S} \vec{p}(\vec{v} \cdot d\vec{A})$$

$$= \hat{x} \int_{S} p_{x} \vec{v} \cdot d\vec{A} + \hat{y} \int_{S} p_{y} \vec{v} \cdot d\vec{A} + \hat{z} \int_{S} p_{z} \vec{v} \cdot d\vec{A}$$

$$= \hat{x} \int_{S} \vec{E}_{1} \cdot d\vec{A} + \hat{y} \int_{S} \vec{E}_{2} \cdot d\vec{A} + \hat{z} \int_{S} \vec{E}_{3} \cdot d\vec{A}$$
(10)

Now we can use the divergence theorem on the three quantities  $\vec{E}_1$ ,  $\vec{E}_2$  and  $\vec{E}_3$  to get

$$\int_{S} \vec{p}(\vec{v} \cdot d\vec{A})$$

$$= \hat{x} \int_{V} \vec{\nabla} \cdot \vec{E}_{1} d^{3}V + \hat{y} \int_{V} \vec{\nabla} \cdot \vec{E}_{2} d^{3}V + \hat{z} \int_{V} \vec{\nabla} \cdot \vec{E}_{3} d^{3}V$$

$$= \hat{x} \int_{V} \vec{\nabla} \cdot (p_{x}\vec{v}) d^{3}V + \hat{y} \int_{V} \vec{\nabla} \cdot (p_{y}\vec{v}) d^{3}V$$

$$+ \hat{z} \int_{V} \vec{\nabla} \cdot (p_{z}\vec{v}) d^{3}V \qquad (11)$$

(A quantity like  $\vec{E}_1 = p_x \vec{v}$  is not really a vector field because it does not transform like a vector when we rotate the coordinate system. But if no rotations are involved,  $p_x$  acts like a scalar field p, and  $p_x \vec{v}$  acts like a vector field  $\vec{j} = p\vec{v}$  in the divergence theorem.)

#### **Einstein Summation Convention**

In Equation (11) we have some fairly mixed up vector components like

$$\hat{\mathbf{x}} \, \vec{\nabla} \cdot (\mathbf{p}_{\mathbf{x}} \mathbf{v}) = \hat{\mathbf{x}} \left[ \nabla_{\mathbf{x}} (\mathbf{p}_{\mathbf{x}} \mathbf{v}_{\mathbf{x}}) + \nabla_{\mathbf{y}} (\mathbf{p}_{\mathbf{x}} \mathbf{v}_{\mathbf{y}}) + \nabla_{\mathbf{z}} (\mathbf{p}_{\mathbf{x}} \mathbf{v}_{\mathbf{z}}) \right]$$
(12)

There is a notation, credited to Einstein, that makes it easy to handle such terms. In Equations (13), we write the dot product of two vectors in three different ways.

$$\vec{a} \cdot \vec{b} = a_x b_x + a_y b_y + a_z b_z$$
(13a)

$$= \sum_{i=x,y,z} a_i b_i \tag{13b}$$

$$= a_i b_i \tag{13c}$$

In (13a) we see the usual definition of the dot product of two vectors. In (13b), we used the index (i) to represent the subscripts x, y, z and included a summation sign to show we are adding up the three terms. Supposedly Einstein got tired of writing summation signs and introduced the notation in (13c). He said that if the index appears twice, then automatically take a sum. As an example, if you encounter  $a_i b_j c_i$ you would sum over the repeated index (i) to get

$$a_i b_j c_i = \sum_{i=x,y,z} a_i b_j c_i$$
  
=  $a_x b_j c_x + a_y b_j c_y + a_z b_j c_z$  (14)

Since the index (j) is not summed over, it remains the same index throughout. We would say that  $a_i b_j c_i$  is the (j)th component of the vector  $a_i \vec{b} c_i$ .

Using this notation in Equation (12), we have

$$\hat{x} \left[ \nabla_{x}(p_{x}v_{x}) + \nabla_{y}(p_{x}v_{y}) + \nabla_{z}(p_{x}v_{z}) \right]$$

$$= \hat{x} \left[ \nabla_{i}(p_{x}v_{i}) \right] = \nabla_{i}(\left[ \hat{x}p_{x} \right]v_{i})$$
(15)

and Equation (11) can be written as

$$\int_{S} \vec{p}(\vec{v} \cdot d\vec{A})$$

$$= \int_{V} \nabla_{i} \left( [\hat{x}p_{x} + \hat{y}p_{y} + \hat{z}p_{z}]v_{i} \right) d^{3}V$$

$$= \int_{V} \nabla_{i}(\vec{p}v_{i}) d^{3}V$$
(16)

Using Equation (16) in Equation (8) gives

$$\frac{\delta \vec{P}_{V}}{\delta t} = \int_{V(t)} \left[ \frac{\partial \vec{p}}{\partial t} + \nabla_{i} (\vec{p} v_{i}) \right] d^{3}V$$
(17)

This is the formula for the rate of change of the momentum of the fluid particles inside the volume V that moves with the particles. It is all expressed in terms of variables at the time (t).

#### **Mass Continuity Equation**

When we substitute  $\vec{p} = \rho \vec{v}$  into Equation (17) we end up with quite a few terms. The result can be simplified by using the equation for the conservation of mass during the flow. The derivation, which is worth repeating, is similar to our derivation in Chapter 10 of the conservation of electric charge.

Consider a volume V bounded by a fixed surface S in a fluid of density  $\rho$ . The rate at which mass is flowing out of V (the mass flux) is given by the integral over S

$$-\frac{\mathrm{d}\mathbf{M}}{\mathrm{d}t} = \int_{S} (\rho \vec{\mathbf{v}}) \cdot \mathrm{d}\vec{\mathbf{A}} \quad \begin{array}{l} \text{rate at which} \\ \text{mass is flowing} \\ \text{out across S} \end{array}$$
(18)

where  $\rho \vec{v}$  is the mass current. We can use the divergence theorem to convert this surface integral to a volume integral, giving

$$-\frac{dM}{dt} = \int_{V} \vec{\nabla} \cdot (\rho \vec{v}) d^{3}V$$
(19)

If mass is flowing out of V, there must be a decrease in the density  $\rho$  inside. The rate at which the total mass inside is decreasing is related to the change in density  $\rho$  by

$$-\frac{\mathrm{d}M}{\mathrm{d}t} = -\int_{\mathrm{V}} \frac{\partial\rho}{\mathrm{d}t} \mathrm{d}^{3}\mathrm{V}$$
(20)

Equating our two formulas for - dM/dt gives

$$\int_{V} \vec{\nabla} \cdot (\rho \vec{v}) d^{3}V = -\int_{V} \frac{\partial \rho}{\partial t} d^{3}V$$
(21)

The two volume integrals can be combined to give

$$\int \left[ \frac{\partial \rho}{\partial t} + \vec{\nabla} \cdot (\rho \vec{v}) \right] d^{3} V = 0$$
 (22)
Since Equation (22) must hold for any volume V or fixed surface S we can construct, the terms in the square brackets must be zero, giving

$$\frac{\partial \rho}{\partial t} + \vec{\nabla} \cdot (\rho \vec{v}) = 0 \qquad \begin{array}{c} mass \ continuity\\ equation \end{array} \tag{23}$$

# Rate of Change of Momentum when Mass is Conserved

With the continuity equation written down, let us return to our formula for the rate of change of the momentum of the fluid particles, replacing the momentum density  $\vec{p}$  by  $\rho \vec{v}$  to get

$$\frac{\delta \vec{P}_{V}}{\delta t} = \int_{V} \left[ \frac{\partial(\rho \vec{v})}{\partial t} + \nabla_{i}(\rho \vec{v} v_{i}) \right] d^{3}V \qquad (24)$$

The terms in the square bracket become

$$\begin{bmatrix} \end{bmatrix} = \frac{\partial \rho}{\partial t} \vec{v} + \rho \frac{\partial \vec{v}}{\partial t} + \vec{v} \nabla_i (\rho v_i) + \rho v_i \nabla_i \vec{v} \\ = \rho \begin{bmatrix} \frac{\partial \vec{v}}{\partial t} + \vec{v}_i \nabla_i \vec{v} \end{bmatrix} + \vec{v} \begin{bmatrix} \frac{\partial \rho}{\partial t} + \vec{\nabla} \cdot (\rho \vec{v}) \end{bmatrix}$$
(25)

where we wrote  $\nabla_i(\rho v_i) = \vec{\nabla} \cdot (\rho \vec{v})$  .

We immediately see that the second bracket is zero by the mass continuity equation, and we are left with our final result

$$\frac{\delta \vec{P}_{V}}{\delta t} = \int_{V} \rho \left[ \frac{\partial v}{\partial t} + (\vec{v} \cdot \vec{\nabla}) \vec{v} \right] d^{3}V$$
(26)

Equation (26) holds even when the density of the fluid is changing.

#### Newton's Second Law

We are now in a position to apply Newton's second law to the fluid in our volume V. Equation (26) gives us the total rate of change of the momentum of the particles within V. We now want to equate that to the total force  $\vec{F}_{tot}$  acting on the particles. We will calculate that by adding up the individual forces per unit volume, which are the pressure force, the viscous force, and the other forces. Then we integrate the sum over the volume V.

In View 3 of Chapter 3 on divergence, we found that the pressure force per unit volume was

$$\vec{\mathbf{f}}_{\mathbf{p}} = -\vec{\nabla}\mathbf{p} \tag{3.3-2}$$

In Chapter 4 we found that the viscous force per unit volume for a constant density Newtonian fluid was

$$\vec{f}_{v} = \mu \nabla^{2} \vec{v} \tag{4-19}$$

Letting  $\vec{f}_{other}$  represent all other forces per unit volume, we get for the total force  $\vec{F}_{tot}$  acting on the fluid within V

$$\vec{F}_{tot} = \int_{V} \left[ -\nabla p + \mu \nabla^2 \vec{v} + \vec{f}_{other} \right] d^2 V \qquad (27)$$

Equating the total force  $\vec{F}_{tot}$  to the rate of change of momentum  $\delta \vec{P}_V / \delta t$ , Equations (27) and (26), gives

$$\vec{F}_{tot} = \frac{\delta \vec{P}_{V}}{\delta t}$$
$$\vec{F}_{tot} = \int_{V} \left[ -\nabla p + \mu \nabla^{2} \vec{v} + \vec{f}_{other} \right] d^{3} V$$
$$= \int_{V} \rho \left[ \frac{\partial \vec{v}}{\partial t} + (\vec{v} \cdot \vec{\nabla}) \vec{v} \right] d^{3} V$$
(28)

Putting everything under a single integral sign gives us

$$\int_{V} \left[ \rho \frac{\partial \vec{v}}{\partial t} + \rho (\vec{v} \cdot \vec{\nabla}) \vec{v} + \vec{\nabla} p - \mu \nabla^{2} \vec{v} - \vec{f}_{other} \right] d^{3} V$$
$$= 0$$
(29)

Next we have our usual argument that Equation (29) must hold for any volume V. The only way we can always get the answer zero for the integral is for the integrand, the stuff in the square brackets, to be zero. Thus we end up with the equation

$$\rho \left[ \frac{\partial \vec{v}}{\partial t} + (\vec{v} \cdot \vec{\nabla}) \vec{v} \right] = - \vec{\nabla} p + \mu \nabla^2 \vec{v} + \vec{f}_{other}$$
(30)

This is one form of the Navier-Stokes equation.

It is usually more convenient to divide through by  $\rho$ , using

$$\nu = \frac{\mu}{\rho} \qquad \begin{array}{c} kinematic \\ viscosity \\ coefficienct \end{array} \qquad (4-41)$$

where v is the so called kinematic viscosity described in the pipe flow experiment of Chapter 4 (page Cal 4-9). We will also define  $\vec{g}_{other}$  by

$$\vec{g}_{other} = \frac{\vec{f}_{other}}{\rho}$$
 other forces  
per unit mass (31)

which represents all other forces, but now as force per unit mass, since we have divided by mass per unit volume  $\rho$ . We get

$$\frac{\partial \vec{v}}{\partial t} + (\vec{v} \cdot \vec{\nabla})\vec{v} = \frac{-\vec{\nabla}p}{\rho} + \nu \nabla^{2}\vec{v} + \vec{g}_{other}$$
*Navier-Stokes Equation*
(32)

Equation (32) is the form of the Navier-Stokes equation you are likely to find in the textbooks. It represents the basic starting point for fluid dynamics theory.

Equation (32) is quite general. Only in the formula  $v\nabla^2 \vec{v}$  for the viscous force have we made any assumptions about the density being constant (i.e.,  $\vec{\nabla} \cdot \vec{v} = 0$ ), and that the coefficient of viscosity v is constant. If we have a non constant density fluid, or non constant coefficient of viscosity, all we have to do is correct the viscosity term.

In Chapter 23 of the Physics text, we began our discussion of vector fields with the velocity field. We made this choice because it is easier to picture a velocity field than an electric field, and we could immediately derive Bernoulli's equation from some simple energy arguments. How things have changed in this chapter! The derivation of the Navier-Stokes equation for the velocity field was harder to do than deriving the wave equations for  $\vec{E}$  and  $\vec{B}$ , and the result is more complex. We have seen terms that resemble  $v\nabla^2 \vec{v}$  and  $\partial \vec{v}/\partial t$  in our discussion of wave equations, but we have not encountered a term that looks anything like  $(\vec{v} \cdot \vec{\nabla})\vec{v}$ .

Not only does  $(\vec{v} \cdot \vec{\nabla})\vec{v}$  have a peculiar combination of components, it is essentially proportional to the square of the velocity field, which makes the Navier-Stokes equation a *non linear* equation. What that means is as follows. The equations we have studied so far, the wave equations for  $\vec{E}$  and  $\vec{B}$ , and Schrödinger's equation for  $\psi$ , are *linear* equations. This means that there are no terms involving the square of  $\vec{E}$ ,  $\vec{B}$  or  $\psi$ , and as a result we have the rule that *waves add*. What this implies is that if you have two solutions to a wave equation. For a non linear equation, the sum of two solutions is also a solution. For a non linear equation.

In the case of water waves, if the amplitudes of the waves are small, the  $(\vec{v} \cdot \vec{\nabla})\vec{v}$  term is not important and waves add, as we saw in the ripple tank experiments. However, if the amplitudes become large, the  $(\vec{v} \cdot \vec{\nabla})\vec{v}$  term, being proportional to  $v^2$ , becomes large and we get non linear effects like the breakers we see when ocean waves come up to the beach. There is no way you can get the solution describing a breaking wave from adding up the solutions for many small amplitude waves. The non linear term brings in completely new physics.

Despite the apparent complexity of the Navier-Stokes equation, some fairly simple results can be derived from it. One is Bernoulli's equation which we will discuss in the next section, the other is a generalized Helmholtz theorem which we will derive after that. In our discussion of Bernoulli's equation we learn more than we did in the Physics text. Here we will determine the conditions when Bernoulli's equation applies, and when it does not.

### **BERNOULLI'S EQUATION**

There is a vector identity which allows us to change the form of the Navier-Stokes equation so that the terms in Bernoulli's equation begin to appear. The vector identity is

$$(\vec{\mathbf{v}} \cdot \vec{\nabla}) \vec{\mathbf{v}} = \vec{\nabla} \left( \frac{\mathbf{v}^2}{2} \right) - \vec{\mathbf{v}} \times (\vec{\nabla} \times \vec{\mathbf{v}})$$
(33)

In Appendix 1 of this chapter we show you a relatively easy way to derive vector identities involving the curl. Equation (33) is the explicit example we use.

Noting that  $\nabla \times \vec{v}$  is the vorticity  $\vec{\omega}$ , we can write Equation (33) as

$$(\vec{\mathbf{v}}\cdot\vec{\nabla})\vec{\mathbf{v}} = \vec{\nabla}\left(\frac{\mathbf{v}^2}{2}\right) - \vec{\mathbf{v}}\times\vec{\boldsymbol{\omega}}$$
 (33a)

Using Equation (33a) for the  $(\vec{v} \cdot \vec{\nabla})\vec{v}$  term in the Navier-Stokes equation (32) gives

$$\frac{\partial \vec{v}}{\partial t} - \vec{v} \times \vec{\omega} = -\vec{\nabla} \left( \frac{v^2}{2} \right) - \frac{\vec{\nabla} p}{\rho} + v \nabla^2 \vec{v} + \vec{g}_{other}$$
(34)

Our next step is to extract the gravitational force from  $\vec{g}_{other}$  and display it explicitly. The gravitational force per unit volume of fluid  $\vec{f}_g$  is

$$\vec{f}_{g} = \rho \vec{g} = \rho(-\vec{\nabla}gy)$$
(35)

where y is the upward directed coordinate and g the acceleration due to gravity. (Take a break and show that  $-\vec{V}(gy)$  is equal to  $\vec{g}$ , a vector of magnitude g pointing down.)

The force terms in Equation (32) are forces per unit mass. We get the gravitational force per unit mass,  $\vec{g}_{gravity}$  by dividing  $\vec{f}_{g}$  by the density  $\rho$ .

$$\vec{g}_{\text{gravity}} = \frac{\vec{f}_g}{\rho} = -\vec{\nabla}(gy)$$
 (36)

The force  $\vec{g}_{other}$  becomes

$$\vec{g}_{other} = -\vec{\nabla}(gy) + \vec{g'}_{other}$$
 (36a)

where  $\vec{g}'_{other}$  represents other forces not including gravity.

Using Equation (36a) in Equation (34) gives

$$\frac{\partial \vec{v}}{\partial t} - \vec{v} \times \vec{\omega}$$

$$= -\vec{\nabla} \left( \frac{v^2}{2} + gy \right) - \frac{\vec{\nabla} p}{\rho} + v \nabla^2 \vec{v} + \vec{g}_{other}^{'}$$
(37)

Up to this point the only place we assumed that  $\rho$  was constant was in the viscosity term  $\nu \nabla^2 \vec{v}$ . But for the remainder of this chapter we will assume that  $\rho$  is constant and use that to simplify other terms. For example, we can pull a constant  $\rho$  inside the gradient, giving

$$-\frac{\vec{\nabla}p}{\rho} = -\vec{\nabla}\left(\frac{p}{\rho}\right) \qquad if \ \rho \ is \\ constant \qquad (38)$$

Using Equation (38) in Equation (37) gives

$$\frac{\partial \vec{v}}{\partial t} - \vec{v} \times \vec{\omega} = -\vec{\nabla} \left( \frac{p}{\rho} + \frac{v^2}{2} + gy \right) + v \nabla^2 \vec{v} + \vec{g'}_{other}$$

constant density fluids

(39)

It is in Equation (39) we see the Bernoulli terms  $(p/\rho + v^2/2 + gy)$ . We can now use the equation both to derive Bernoulli's equation and to state the conditions under which it applies.

Suppose we have the following four conditions: (1) constant density, (2) a steady flow so that  $\partial \vec{v} / \partial t = 0$ , (3) that viscosity is not important so that we can neglect the viscosity term  $v\nabla^2 \vec{v}$ , and (4) that there are no forces other than pressure and gravity acting on the fluid so that we can set  $\vec{g'}_{other} = 0$ . These conditions are

$$\rho = \text{constant}$$

$$\frac{\partial \vec{v}}{\partial t} = 0 \quad \text{steady flow}$$

$$\nu \nabla^2 \vec{v} = 0 \quad \text{neglect viscosity}$$

$$\vec{g'}_{\text{other}} = 0 \quad \text{no other forces}$$
(40)

Under conditions (40) the Navier-Stokes equation becomes

$$\vec{v} \times \vec{\omega} = \vec{\nabla} \left( \frac{p}{\rho} + \frac{v^2}{2} + gy \right)$$
 (41)

## **Applies Along a Streamline**

In Chapter 23 of the Physics text, we called the collection of Bernoulli terms the *hydrodynamic voltage*. Labeling their sum by  $\phi_H$ , we have

$$\phi_{\rm H} \equiv \frac{p}{\rho} + \frac{v^2}{2} + gy \qquad {hydrodynamic} \\ voltage \qquad (42)$$

With this notation, Equation (41) becomes

$$\vec{\mathbf{v}} \times \vec{\boldsymbol{\omega}} = \vec{\nabla} \boldsymbol{\phi}_{\mathrm{H}} \tag{43}$$

We used the name hydrodynamic voltage for  $\phi_H$  to stress the similarity between hydrodynamic voltage-drops in a fluid circuit and electric voltage-drops in an electric circuit.

Later in the Physics text, in our discussion of electric voltage in Chapter 25, we changed the name from voltage to potential, and started constructing contour maps of the potential  $\phi$ . Our main example was the map of the electric potential produced by charges +3 and -1 shown in Figure (25-15) reproduced again here. The lines of constant potential are the contour lines, and the lines of steepest descent are the field lines.

In our discussion of gradient in this text, we saw that the gradient vector  $\vec{\nabla}\phi$  pointed along the field lines. Or to say it another way, the gradient  $\vec{\nabla}\phi$  was a maximum in the direction where the slope is the steepest, and was zero in the direction of a contour line where the value of  $\phi$  remains constant.

Our Equation (43),  $\vec{v} \times \vec{\omega} = \vec{\nabla} \phi_H$ , is an equation relating the gradient of the potential  $\phi_H$  to what at first looks like a rather complicated term  $\vec{v} \times \vec{\omega} = \vec{v} \times (\vec{\nabla} \times \vec{v})$ . But there is one thing that is simple about  $\vec{v} \times \vec{\omega}$ . Because of the cross product,  $\vec{v} \times \vec{\omega}$  is always perpendicular to  $\vec{v}$ , i.e., always zero in the direction of  $\vec{v}$ . In a fluid flow, the streamlines follow in the direction of the velocity field  $\vec{v}$ . Thus if we move in the direction of a streamline, we are moving in a direction where  $\vec{v} \times \vec{\omega}$  and thus  $\vec{V} \phi_H$  is zero. But if we move in a direction where the gradient of  $\phi_H$  is zero, we must be moving along a contour line of  $\phi_H$ , and the value of  $\phi_H$  must be constant. Thus the physical content of the equation  $\vec{V} \phi_H = \vec{v} \times \vec{\omega}$  is that  $\phi_H$  *is constant along a streamline*. Re-expressing  $\phi_H$  as  $p/\rho + v^2/2 + gy$ , we get the result

$$\frac{p}{\rho} + \frac{v^2}{2} + gy = \frac{\text{constant along}}{\text{a streamline}}$$
(44)

when conditions (40) are obeyed.

Equation (44), with the associated conditions, is our precise statement of Bernoulli's equation. It tells us both when Bernoulli's equation can be used, and why it should be applied along a streamline. In the special case of potential flow where  $\vec{\omega} = \vec{\nabla} \times \vec{v}$  is zero everywhere, then Equation (41) becomes  $\vec{\nabla}\phi_H = 0$ , which implies  $\vec{\nabla}\phi_H = p/\rho + v^2/2 + gy = constant$  throughout the fluid. For potential flow we do not have to apply Bernoulli's equation only along a streamline.



**Figure 25-15 (repeated)** The lines of equal height, the contour lines, are the lines along which the potential  $\phi$  is constant.

### The Viscosity Term

Although the Navier-Stokes equation is a rather formidable equation, we are beginning to see some fairly simple or recognizable results emerge. A lot can be learned by studying the nature of the terms in the equation. Here we will see that the viscous force term  $v\nabla^2 \vec{v}$  can be re-expressed in a form that gives one a better understanding of the nature of vortices.

Back in Chapter 8 on the curl, we proved the vector identity

$$\vec{\nabla} \times (\vec{\nabla} \times \vec{A}) = -\nabla^2 \vec{A} + \vec{\nabla} (\vec{\nabla} \cdot \vec{A})$$
(8-5)

If we apply this to the velocity field  $\vec{v}$  of a constant density fluid where  $\vec{\nabla} \cdot \vec{v} = 0$ , we get

$$\nabla^2 \vec{v} = -\vec{\nabla} \times (\vec{\nabla} \times \vec{v}) = -\vec{\nabla} \times \vec{\omega}$$
(45)

Where  $\vec{\omega} = \vec{\nabla} \times \vec{v}$ . Thus the viscous force term in the Navier-Stokes equation can be written as

$$\nu \nabla^2 \vec{v} = -\nu \vec{\nabla} \times \vec{\omega} \qquad viscous \ force \\ per \ unit \ mass \qquad (46)$$

From Equation (46) we see that there are no viscous forces where the vorticity  $\vec{\omega}$  is zero, or even when  $\vec{\omega}$  is constant as in solid body rotation.

In our discussion of vortices in the last chapter, we pictured an ideal vortex as one whose velocity field  $\vec{v}$  was analogous to the magnetic field of a current in a straight wire. If the current in the wire is uniform, then  $\vec{\nabla} \times \vec{B} = \mu_0 \vec{i}$  is a constant inside the wire and zero outside. Thus in our ideal vortex,  $\vec{\omega} = \vec{\nabla} \times \vec{v}$  is uniform inside the core (representing a solid body rotation of the fluid there), and  $\vec{\omega} = 0$  outside where we have the  $\hat{\theta}$  directed 1/r velocity field.

With our new formula  $-v\vec{\nabla}\times\vec{\omega}$  for the viscous force, we see that there is no viscous force acting inside the core where  $\vec{\omega} = \text{constant}$ . What is surprising is that there is also no viscous force acting outside the core in the 1/r circular velocity field. The only place where viscous forces act in an ideal vortex is at the boundary between the core and the fluid outside. The fact that viscous forces do not act either inside or outside the core of an ideal vortex is one reason for the permanence of the vortex structure.

Because the velocity field of a vortex ring is analogous to the magnetic field of a current loop, the fact that  $\vec{\nabla} \times \vec{B} = \mu_0 \vec{i}$  is zero outside the wire loop, implies that the vorticity  $\vec{\omega} = \vec{\nabla} \times \vec{v}$  is zero outside the core of a vortex ring. Thus in a vortex ring or a smoke ring, viscous forces do not act on the fluid outside the core.

 $\vec{F}_{magnus} = -\rho \vec{V}_{rel} \times \vec{\kappa}$  acting on that vortex. But there is no extra mass associated with a fluid core vortex, so one must treat the vortex as a massless object, with the result that the net force on the vortex must be zero. That means that there must be an external force  $\vec{F}_{external}$  acting on the vortex to cancel the Magnus lift force. That is, one must have

$$\vec{F}_{external} + \vec{F}_{magnus} = 0$$
 (108)

### THE HELMHOLTZ THEOREM

While Bernoulli's theorem may be the most famous theorem of fluid dynamics, Helmholtz's theorem is perhaps the most dramatic. To see a smoke ring emerge from an amorphous cloud of smoke and travel across a room in a straight line has to be one of the impressive phenomena of physics. Yet we saw that it was explained by Helmholtz's theorem that in the absence of non potential forces, the fluid particles become trapped on, and move with, the vortex lines.

In this section we will derive Helmholtz's theorem from the Navier-Stokes equation. As a result, all the phenomena we have seen that are explained by Helmholtz's theorem can be viewed as being a consequence of the Navier-Stokes equation.

#### **Equation for Vorticity**

The first step in deriving Helmholtz's theorem is to turn the Navier-Stokes equation into an equation for the vorticity field  $\vec{\omega}$ . We do this by taking the curl of both sides of Equation (39). We have

$$\vec{\nabla} \times \left[ \frac{\partial \vec{v}}{\partial t} - \vec{v} \times \vec{\omega} \right]$$
$$= \vec{\nabla} \times \left[ -\vec{\nabla} \left( \frac{p}{\rho} + \frac{v^2}{2} + gy \right) - v \vec{\nabla} \times \vec{\omega} + \vec{g'}_{other} \right]$$
(47)

where we used Equation (46) to replace  $\nu \nabla^2 \vec{v}$  by  $-\nu \vec{\nabla} \times \vec{\omega}$ .

At this point you might be discouraged by the number of cross products that appear in Equation (47). But immediately there is noticeable simplification. Recall that the curl of a gradient is identically zero,

$$\vec{\nabla} \times \vec{\nabla} \phi \equiv 0 \qquad any \phi \tag{48}$$

Thus the Bernoulli terms all go out in Equation (47)

$$\vec{\nabla} \times \vec{\nabla} \left( \frac{p}{\rho} + \frac{v^2}{2} + gy \right) = 0$$
(49)

which considerably shortens the equation.

Next, we note that because we can interchange the order of partial differentiation, we get

$$\vec{\nabla} \times \frac{\partial \vec{v}}{\partial t} = \frac{\partial}{\partial t} (\vec{\nabla} \times \vec{v}) = \frac{\partial \vec{\omega}}{\partial t}$$
(50)

Thus Equation (47), the curl of the Navier-Stokes equation, becomes

$$\frac{\partial \vec{\omega}}{\partial t} - \vec{\nabla} \times (\vec{v} \times \vec{\omega}) = \vec{\nabla} \times \vec{g}$$
(51)

where  $\vec{g}$ , given by

$$\vec{g} = -\nu \vec{\nabla} \times \vec{\omega} + \vec{g'}_{other}$$
 (52)

represents all forces per unit mass acting on the fluid, except pressure and gravity. Equation (51) is the differential equation for the dynamical behavior of the vorticity field  $\vec{\omega}$ . The only restriction is that it applies to constant density fluids. If we wish to work with non constant density fluids we have to go back and work with Equation (39) and perhaps use a more general formula for the viscous force.

#### **Non Potential Forces**

An important simplification we obtained in going to an equation for the vorticity field  $\vec{\omega}$  was the elimination of the Bernoulli terms. This removes the pressure and gravitational forces from the equation for  $\vec{\omega}$ , implying that pressure and gravity have no direct effect on the behavior of vorticity. We saw this result in the case of the motion of a smoke ring. The ring moved in a straight line across the room completely unaffected by gravity. (Pressure and gravity can have an indirect effect in that they affect the velocity field  $\vec{v}$  which appears in the  $\vec{\nabla} \times (\vec{v} \times \vec{\omega})$ term.) In Equation (51),

$$\frac{\partial \vec{\omega}}{\partial t} - \vec{\nabla} \times (\vec{v} \times \vec{\omega}) = \vec{\nabla} \times \vec{g} \qquad (51) \text{ repeated}$$

the only force terms that survive are those with a non zero curl like the viscosity term. Let us introduce the terminology *potential force*  $\vec{g}_{\phi}$  and a *non potential force*  $\vec{g}_{np}$ . Potential forces are those that can be expressed as the gradient of a potential  $\phi$ , and thus have a zero curl

$$\vec{g}_{\phi} = -\vec{\nabla}\phi \quad ; \quad \vec{\nabla} \times \vec{g}_{\phi} = 0$$
 (53)

while non potential forces  $\vec{g}_{np}$  have non zero curl

$$\vec{\nabla} \times \vec{g}_{np} \neq 0 \tag{54}$$

and thus survive the curl in Equation (51). As a result we can write Equation (51) in the form

$$\frac{\partial \vec{\omega}}{\partial t} - \vec{\nabla} \times (\vec{v} \times \vec{\omega}) = \vec{\nabla} \times \vec{g}_{np} \qquad \begin{array}{c} \text{vortex} \\ \text{dynamics} \\ \text{equation} \end{array} (55)$$

We will call Equation (55) the vortex dynamics equation.

To be quite general, one might like to separate an arbitrary force field  $\vec{g}$  into its potential part  $\vec{g}_{\phi}$  and its non potential part  $\vec{g}_{np}$ , writing

$$\vec{g} = \vec{g}_{\phi} + \vec{g}_{np} \tag{56}$$

The problem is that there is no unique separation of an arbitrary vector field into potential and non potential parts. The only thing that is unique is the curl

$$\vec{\nabla} \times \vec{g} = \vec{\nabla} \times \vec{g}_{np} \tag{57}$$

Physically, Equation (57) is telling us that if we accidentally included some potential terms in our formula for  $\vec{g}_{np}$ , they would disappear when we took the curl in Equation (57).

For a practical matter, the best thing to do is to include all obviously potential forces like pressure and gravity in  $\vec{g}_{\phi}$ , and leave all others that are not obviously potential forces, like the viscous force  $-\nu \vec{\nabla} \times \vec{\omega}$ , in the non potential category  $\vec{g}_{np}$ .

#### A VECTOR IDENTITY FOR A MOVING CIRCUIT

Before we obtain a really clear interpretation of the vortex dynamics equation (55), we need a way of understanding the impact of the rather complex looking term  $-\vec{\nabla} \times (\vec{v} \times \vec{\omega})$ . In this section, we will derive a vector identity that will lead to a strikingly simple interpretation of the combination of terms  $\partial \vec{\omega} / \partial t - \vec{\nabla} \times (\vec{v} \times \vec{\omega})$ . The vector identity involves the rate of change of flux of a solenoidal field like  $\vec{\omega}$  through a circuit that moves with the fluid particles.

It takes a considerable effort to derive this vector identity, an effort involving steps somewhat similar to those we used to calculate the rate of linear momentum in a moving volume. But the resulting simplification in the interpretation of the vortex dynamics equation is more than worth the effort.

To emphasize the general nature of the vector identity, we will calculate the rate of change of the flux of a vector field  $\vec{A}$  through the circuit C' that moves with the fluid particles. The restriction on  $\vec{A}$  will be that it is a solenoidal field with  $\vec{\nabla} \cdot \vec{A} = 0$ .

Let the circuit C'(t) shown in Figure (4) be attached to the fluid particles through which it passes. As time progresses from (t) to  $(t + \delta t)$ , the fluid motion will carry the circuit from position C'(t) to the position C'(t +  $\delta t$ ) as shown. We will also assume that there is a divergence free vector field  $\vec{A}(t)$  in the fluid at time (t). At time (t +  $\delta t$ ) the vector field will

have changed to  $\vec{A}(t + \delta t)$ . What we wish to calculate is the change in the flux of  $\vec{A}$  through the circuit C' as we go from (t) to  $(t + \delta t)$ . We will do the



**Figure 4** *The circuit C' moves with the fluid particles.* 

calculation throwing out terms of order  $\delta t^2$  compared to  $\delta t\,.$ 

At time t, the flux  $\Phi(t)$  of  $\vec{A}$  through C'(t) is

$$\Phi(t) = \int_{\substack{S'(t)}} \vec{A}(t) \cdot d\vec{S}$$
(58)

where S' is a surface bounded by C'(t). At time  $(t + \delta t)$  the flux has become

$$\Phi(t+\delta t) = \int_{\substack{S'(t+\delta t)}} \vec{A}(t+\delta t) \cdot d\vec{S}$$
(59)

The change in flux  $\delta \Phi$  during the time  $\delta t$  is

$$\delta \Phi = \int_{S'(t+\delta t)} \vec{A}(t+\delta t) \cdot d\vec{S} - \int_{S'(t)} \vec{A}(t) \cdot d\vec{S} \qquad (60)$$

Using a Taylor series expansion we can write

$$\vec{A}(t + \delta t) = \vec{A}(t) + \frac{\partial A}{\partial t} \delta t + 0(\delta t^2)$$
 (61)

Thus

$$\delta \Phi = \int_{S'(t+\delta t)} \vec{A}(t) \cdot d\vec{S} - \int_{S'(t)} \vec{A}(t) \cdot d\vec{S} + \delta t \int_{S'(t+\delta t)} \frac{\partial \vec{A}}{\partial t} \cdot d\vec{S}$$
(62)

To calculate the effect of the first two terms in Equation (62), consider the guitar shaped volume shown in Figure (5). The top of the volume is bounded by the curve  $C'(t + \delta t)$ , while the bottom by C'(t). A certain amount of flux  $\Phi_1$ 

$$\Phi_{1} = \int_{\substack{\mathbf{S}'(\mathbf{t})}} \vec{\mathbf{A}}(\mathbf{t}) \cdot \mathbf{d}\vec{\mathbf{S}}$$
(63)

enters up through the bottom of the volume. Some more flux,  $\Phi_2$  flows in through the sides, and an amount  $\Phi_3$ 

$$\Phi_3 = \int_{\substack{S'(t+\delta t)}} \vec{A}(t) \cdot d\vec{S}$$
((64)

flows out through the top.

Because  $\vec{A}(t)$  is a divergence free field [ $\vec{\nabla} \cdot \vec{A}(t) = 0$ ], all the flux flowing in through the bottom,  $\Phi_1$ , and the sides,  $\Phi_2$ , must flow out through the top,  $\Phi_3$ , giving

$$\Phi_3 = \Phi_1 + \Phi_2 \tag{65}$$

(Any of these fluxes could be negative, indicating  $\vec{A}$  pointing in other directions, but all signs are correctly handled by the formalism.)

Using Equations (63) and (64), our formula (62) for  $\delta\Phi$  becomes

$$\delta \Phi = \Phi_3 - \Phi_1 + \delta t \int_{\substack{\substack{\partial \vec{A} \\ \partial t}}} \frac{\partial \vec{A}}{\partial t} \cdot d\vec{S}$$

With  $\Phi_3 = \Phi_1 + \Phi_2$  we get

$$\delta \Phi = \Phi_2 + \delta t \int_{\substack{\mathbf{X}'(t+\delta t)}} \frac{\partial \vec{\mathbf{A}}}{\partial t} \cdot d\vec{\mathbf{S}}$$
(66)

Equation (66) tells us that the change in the flux of  $\vec{A}(t)$  through the moving circuit C'(t) is made of two parts. One is due to the change  $\partial \vec{A}(t)/\partial t$  of the field itself, the other to flux coming in from the sides.



#### Figure 5

Volume bounded by the curves  $C'(t + \delta t)$  and C'(t). The drawing shows flux entering through the bottom and sides, and flowing out through the top.

Our problem now is to calculate the flux  $\Phi_2$  flowing in through the sides of our volume shown in Figure (5). The calculation of  $\Phi_2$  turns out not to be so hard. In Figure (6) we show a small piece of the side of our volume. A fluid particle that is located at position (1) in that diagram at time (t), moves to position (2) during the time  $\delta t$ . The distance from (1) to (2) is described by the displacement vector  $\vec{v}\delta t$  as shown.

We also mark a short length  $d\vec{\ell}$  of the path C'(t) starting at position (1). If we take the cross product of  $\vec{v}\delta t$  with  $d\vec{\ell}$ , we get a vector  $d\vec{S}$  that points into the volume, perpendicular to both  $\vec{v}\delta t$  and  $d\vec{\ell}$ . The length of  $d\vec{S}$  is equal to the area of the parallelogram defined by  $\vec{v}\delta t$  and  $d\vec{\ell}$ . Thus  $d\vec{S}$  represents the inward area vector for the shaded area in Figure (6). The flux  $d\Phi_2$  of  $\vec{A}(t)$  in through this side area  $d\vec{S}$  is

$$d\Phi_{2} = \vec{A}(t) \cdot d\vec{S} = \vec{A}(t) \cdot [(\vec{v}\delta t) \times d\vec{\ell}]$$
  
=  $\delta t \left[ \vec{A}(t) \cdot (\vec{v} \times d\vec{\ell}) \right]$  (67)

In the appendix to this chapter, where we show you an easy way to handle vector identities involving cross products, we derive the identity

$$\vec{A} \cdot (\vec{B} \times \vec{C}) = (\vec{A} \times \vec{B}) \cdot \vec{C}$$
(68)

Using this identity, we can write Equation (67) in the form

$$d\Phi_2 = \delta t[\vec{A}(t) \times \vec{v}] \cdot d\vec{\ell}$$
(69)



The area element ds on the side of our volume.

To calculate the total flux  $\Phi_2$  in through the sides of our volume, all we have to do is integrate the contributions  $\Phi_2$  around the circuit C'(t).We get

$$\Phi_{2} = \delta t \oint_{\substack{C'(t) \\ C'(t)}} [\vec{A}(t) \times \vec{v}] \cdot d\vec{\ell}$$
(70)

Stokes' law, derived in Chapter 8 relates the integral of a vector field  $\vec{B}$  around a closed path to the flux of  $\vec{\nabla} \times \vec{B}$  through the path. We had

$$\oint_{C} \vec{B} \cdot d\vec{\ell} = \int_{S} \vec{\nabla} \times \vec{B} \cdot d\vec{S}$$
(8-31)

where S is the surface bounded by the closed curve C. If we set  $\vec{B} = \vec{A}(t) \times \vec{v}$ , C = C'(t) and S = S'(t), Equation (8-31) becomes

$$\oint_{C'(t)} \begin{bmatrix} \vec{A}(t) \times \vec{v} \end{bmatrix} \cdot d\vec{\ell} = \int_{S'(t)} \vec{\nabla} \times \begin{bmatrix} \vec{A}(t) \times \vec{v} \end{bmatrix} \cdot d\vec{S}$$
(71)

As a result, the flux  $\Phi_2$  of  $\vec{A}(t)$  flowing in through the sides of our volume is

$$\Phi_{2} = \delta t \int_{\mathbf{S}'(t)} \vec{\nabla} \times \left[ \vec{A}(t) \times \vec{v} \right] \cdot d\vec{S}$$
(72)

Using this result in Equation (66) for the change in flux  $\delta \phi$  through our moving circuit gives

$$\delta \Phi = \Phi_2 + \delta t \int_{\substack{S'(t+\delta t)\\S'(t)}} \frac{\partial \vec{A}}{\partial t} \cdot d\vec{S}$$
(66) repeated  
$$= \delta t \int_{\substack{S'(t)\\S'(t)}} \vec{\nabla} \times \left[ \vec{A}(t) \times \vec{v} \right] \cdot d\vec{S} + \delta t \int_{\substack{S'(t+\delta t)\\S'(t+\delta t)}} \frac{\partial \vec{A}}{\partial t} \cdot d\vec{S}$$
(73)

At this point everything is evaluated at the time (t) except for the integral of the flux of  $\partial \vec{A}(t)/\partial t$  at the surface  $S'(t + \delta t)$ .

As we have just seen, the flux of any vector field through  $S'(t + \delta t)$  is equal to the flux through the end S'(t) plus a term like  $\Phi_2$  representing a flow in through the sides. Because the flux in through the sides is of the order  $\delta t$  smaller than the flow in through the end, and because the  $\partial \vec{A}/\partial t$  term already has a factor of  $\delta t$ , our neglect of the flux of  $\partial \vec{A}/\partial t$  in through the sides will be an error of order  $\delta t^2$  which may be ignored. Thus we can replace  $S'(t + \delta t)$  by S'(t) in Equation (73). Dividing through by  $\delta t$ , and for later convenience replacing  $\vec{A}(t) \times \vec{v}$  by  $-\vec{v} \times \vec{A}(t)$ , we get

$$\frac{\delta\Phi(\vec{A})}{\delta t} = \int_{S'(t)} \left\{ \frac{\partial \vec{A}(t)}{\partial t} - \vec{\nabla} \times \left[ \vec{v} \times \vec{A}(t) \right] \right\} \cdot d\vec{S}$$
(74)

Equation (74) is the general formula for the rate of change of flux of the vector  $\vec{A}(t)$  through a circuit C'(t) that moves with the fluid particles. The circuit C'(t) bounds the surface S'(t), and it is assumed that  $\vec{A}$  is a solenoidal field ( $\vec{\nabla} \cdot \vec{A} = 0$ ).

# The Integral Form of the Vortex Dynamics Equation

Although the derivation of Equation (74) was rather lengthy, the result can be immediately applied to our vortex dynamics Equation (55). If we integrate Equation (55) over a surface S'(t) bounded by a circuit C'(t) we get

$$\int \left[ \frac{\partial \vec{\omega}(t)}{\partial t} - \vec{\nabla} \times (\vec{v} \times \vec{\omega}) \right] \cdot d\vec{S} = \int [\vec{\nabla} \times \vec{g}_{np}] \cdot d\vec{S}$$
  
S'(t) (75)

Because the vorticity  $\vec{\omega}$  is always a solenoidal field, we can replace  $\vec{A}(t)$  by  $\vec{\omega}(t)$  in Equation (74) and immediately recognize the left side of Equation (75) as the rate of change of the flux of  $\vec{\omega}$  through the moving circuit C'(t). Calling this rate  $\delta \Phi(\vec{\omega})/\delta t$ , we have

$$\int \left[ \frac{\partial \vec{\omega}(t)}{\partial t} - \vec{\nabla} \times (\vec{v} \times \vec{\omega}) \right] \cdot d\vec{S} = \frac{\delta \Phi(\vec{\omega})}{\delta t}$$
(76)

On the right side of Equation (75), we can use Stokes' theorem to replace the surface integral of  $\vec{\nabla} \times \vec{g}_{np}$  over S'(t) by the line integral of  $\vec{g}_{np}$ around C'(t) giving

$$\int_{S'(t)} [\vec{\nabla} \times \vec{g}_{np}] \cdot d\vec{S} = \oint_{C'(t)} \vec{g}_{np} \cdot d\vec{\ell}$$
(77)

Combining Equations (76) and (77) gives us the general vortex dynamics Equation (78), a result which assumes only that  $\rho$  is constant.

the rate of change of the  
flux of 
$$\vec{\omega}$$
 through a  
circuit C'(t) that moves  
with the fluid particles  $\frac{\delta \Phi_{\omega}}{\delta t} = \oint_{C'(t)} \vec{g}_{np} \cdot d\vec{\ell}$ 

extended Helmholtz equation (78)

It seems rather remarkable that an equation as complex looking as the Navier-Stokes equation can be converted, by taking the curl, to something simple enough to be described almost completely in words. In a sense the only calculation we have to do to apply Equation (78), is to calculate the line integral of a non potential force  $\vec{g}_{np}$  around a closed path. For reasons that will become clear shortly, we will call Equation (78) the *extended Helmholtz equation*.

# The Helmholtz Theorem

It is an immediate step to go from Equation (78) to Helmholtz's famous theorem of 1858. If there are no non potential forces acting on the fluid, i.e., if  $\vec{g}_{np} = 0$ , then we get the simple statement

If there are no non potential forces acting on the fluid, then there is no change in the flux of  $\overline{\omega}$  through any closed circuit C'(t) that moves with the fluid particles

Helmholtz theorem (79)

At this point we have reduced much of fluid dynamics to a simple word equation.

Equation (79) is perhaps the most precise statement of Helmholtz's theorem, but equivalent statements are also enlightening. Suppose, for example, we define a *vortex line* as a *small unit flux tube of*  $\vec{\omega}$ . Because  $\vec{\omega}$  is solenoidal, the flux tubes or vortex lines cannot stop or start in the fluid. Equation (79) tells us that, in the absence of non potential forces, the number of vortex lines threading any circuit C'(t), i.e., the total flux of  $\vec{\omega}$ , remains constant as the circuit moves with the fluid particles. This clearly will happen if the lines themselves move with the fluid.

Equation (79) does not actually require, in all cases, that the vortex lines must move with the fluid particles. As we saw back in Chapter 12, the vorticity  $\vec{\omega}$  is uniform for solid body rotation. Thus the flux of  $\vec{\omega}$  will remain constant through any circuit C'(t) moving with the fluid, whether or not we think of the vortex lines themselves as moving with the fluid. With a uniform  $\vec{\omega}$ , we cannot tell if the vortex lines are moving or not.

We saw, however, that the situation is very different when dealing with a quantum fluid where the vorticity  $\vec{\omega}$ , although roughly imitating solid body rotation, is lumped up in the vortex cores. In this case Equation (79) clearly requires that the separate vortex cores move around with the fluid. We can easily tell whether lumped up vorticity is moving.

There is, however, no harm in assuming that the vortex lines move with the fluid for solid body rotation. This interpretation has the advantage that if a slight perturbation is introduced into the vorticity field, we can follow the perturbation and see that the associated lines do move.

# EXTENDED HELMHOLTZ THEOREM

If the Helmholtz theorem tells us that in the absence of non potential forces, vortex lines move with the fluid particles, then what happens when non potential forces are present? What is the effect on vorticity of a force  $\vec{g}_{np} \neq 0$ ? The answer, which we obtain from our vortex dynamics Equation (78) is quite simple. It is that *the non potential forces*  $\vec{g}_{np}$  cause a relative motion of the vortex lines and the fluid particles.

It was the study of the behavior of quantized vortices in superfluid helium and superconductors that led to a more complete understanding of the effect of non potential forces on vortex motion. One experiment in particular, an experiment by Rayfield and Reif involving charged vortex rings in superfluid helium, is what initiated this detailed study. We will use a discussion of the Rayfield-Reif experiment to develop the ideas contained in the extended Helmholtz theorem.

# **The Rayfield-Reif Experiment**

Rayfield and Reif were able to create their charged vortex rings by placing a radioactive substance in a container of superfluid helium. The radioactive substance emitted charged particles, either electrons or protons, depending on the substance. What they found was that the charged particle, moving through the superfluid, would create quantized vortex rings in the superfluid, and then in a process still not perfectly understood, the charged particle would become trapped in the core of the ring it created, producing an electrically charged vortex ring.

The interesting part about having an electrically charged vortex ring, is that you can apply an electric field and exert an electric force on the core of the ring. We will see that this electric force acting on the core represents a non potential force acting on the fluid in the region of the core. As a result, Rayfield and Reif were able to study, in detail, the effects of non potential forces acting on vortex lines. Their experiments provided a superb verification of Equation (78) and the interpretation that non potential forces cause a relative motion of the vortex lines and the fluid particles. To apply Equation (78) to the Rayfield-Reif experiment, consider Figure (7) where we show the cross section of a vortex core with a force density  $\vec{g}$  acting on the fluid in the core. The force  $\vec{g}$  represents the electric force acting on the charged fluid in the core. Outside the core there is no force where the fluid is electrically neutral.

On Figure (7) we have drawn three contours labeled  $C'_1$ ,  $C'_2$ , and  $C'_3$ . The primes indicate that these paths are moving with the fluid particles, and that we are looking at the paths now at time (t). If we integrate  $\vec{g}$  around contour  $C'_1$ , we get a positive contribution along the bottom section of the path, and no contribution from the other sections that lie outside the core. Thus we get

$$\oint_{C'_1} \vec{g} \cdot d\vec{\ell} = \text{positive number}$$
(80)

For the force density  $\vec{g}$  to be a conservative potential force, we would have to have  $\oint \vec{g} \cdot d\vec{\ell} = 0$  for any possible path. Because the integral is not zero for circuit C'<sub>1</sub>, Equation (80) shows that  $\vec{g}$  is a non potential force.

To see what a localized force like  $\vec{g}$  cannot do, look at the path C'<sub>3</sub> that goes completely around the core and lies completely in a region where  $\vec{g} = 0$ . For this path we get

$$\oint_{C'_3} \vec{g} \cdot d\vec{\ell} = 0$$
(81)

Thus from Equation (78) we find that there is no change in the flux of  $\vec{\omega}$  through the path C'<sub>3</sub>. Since C'<sub>3</sub> goes around the entire core, the flux of  $\vec{\omega}$  through C'<sub>3</sub> is the total circulation k of the vortex. Thus a *localized non potential force*, (one where we can draw a circuit like C'<sub>3</sub> that is in the fluid but outside the force) *cannot change the circulation*  $\kappa$  *of the vortex line*.

If  $\vec{g}$  cannot change the circulation  $\kappa$ , what does it do? To find out we look more closely at the paths  $C'_1$  and  $C'_2$  lying above and below the line. We saw in Equation (80) that  $\oint \vec{g} \cdot d\vec{\ell}$  was a positive number for the upper path  $C'_1$ . Thus  $\vec{g}$  must be causing an increase in the flux of  $\vec{\omega}$  through the upper path. When we integrate  $\vec{g}$  around the lower path C'<sub>2</sub>, we get zero except where the path comes back through the core, in a direction opposite to  $\vec{g}$ , making  $\oint \vec{g} \cdot d\vec{\ell}$  negative there. As a result

$$\oint_{C'_2} \vec{g} \cdot d\vec{\ell} = \text{negative number}$$
(82)

and we find that  $\vec{g}$  is causing a decrease in the flux of  $\vec{\omega}$  through the lower path.

What does it mean when we see that  $\vec{g}$  is causing the flux of  $\vec{\omega}$  to decrease in the lower path, increase in the upper path, but not change the total flux of the core? It means that  $\vec{g}$  is *causing the vortex line to move upward*. Since the paths C'<sub>1</sub> and C'<sub>2</sub> are attached to the fluid particles, the flow of  $\vec{\omega}$  from the lower path to the upper path represents an upward motion of the vortex line *relative to the fluid particles*. Thus the non potential force  $\vec{g}$  causes a relative motion of the vortex lines and the fluid particles, a relative motion that is absent if there are no non potential forces acting on the fluid.



#### Figure 7

An external force  $\vec{g}$  is applied to the fluid in the core of a vortex. We see that the  $( \oint \vec{g} \cdot d^{\vec{\ell}} )$  is positive around the upper path  $C'_1$ , meaning that flux of  $\vec{\omega}$  is increasing through that path. The integral is negative through the lower path  $C'_2$  meaning that flux of  $\vec{\omega}$  is decreasing there. This results in an upward flow of vorticity. Since  $( \oint \vec{g} \cdot d^{\vec{\ell}} = 0)$  for the big path surrounding the entire core, the total flux, or total circulation  $\kappa$ , is unchanged. This relative motion of the vortex line is sketched in Figure (8), where we designate the relative velocity by the vector  $\vec{v}_{rel}$ . Note that the motion is gyroscope like; when we push in the  $\hat{x}$  direction on a  $\hat{z}$  oriented vortex line, the line moves, not in the direction we push, but up in the  $\hat{y}$  direction.

# Exercise 1

Use Equation (78) and Figure (9) to show that the vortex line has no relative velocity in the direction that  $\vec{g}$  pushes on the fluid.

# Exercise 2

What is the direction of the relative velocity  $\vec{v}_{rel}$  if  $\vec{g}$  is  $\hat{x}$  directed as in Figure (8), but  $\vec{\omega}$  points in the  $-\hat{z}$  direction? (I.e., what happens if we reverse  $\vec{\omega}$ ?) Explain using Equation (78).



Figure 8

The relative velocity  $\vec{v}_{rel}$  of the vortex caused by the non potential force  $\vec{g}$ .



Figure 9

Paths for determining the relative motion of the line in the direction of the force  $\vec{g}$ .

# **Motion of Charged Vortex Rings**

Now that we have some idea of the effect of a localized force acting on a vortex line, let us return to our discussion of the Rayfield-Reif experiment.

As we mentioned, Rayfield and Reif created charged vortex rings in superfluid helium by placing a radioactive substance in the superfluid that emitted charged particles, either an electron or a proton depending on the substance. They ended up with charged objects in the superfluid, objects whose motion they could control using electric fields, and whose speed they could measure by timing a pulse of the particles moving between two grids.

But how could they know that the charged objects in the superfluid were actually vortex rings? The objects were tiny, carrying the charge of only one proton or one electron. In addition the core of a quantum vortex is of the order of an atomic diameter, so that the rings they were dealing with could be as small as only a few tens of atomic diameters. How could they be sure that these objects, that were much too small to be seen, were actually vortex rings?

The answer was in the peculiar behavior of these objects, a behavior only exhibited by vortex rings. The more they accelerated these objects, *the harder they pushed on them,* with an electric field, *the slower they went!* The reason for this behavior follows directly from the extended Helmholtz equation, Equation (78).

In Figure (10) we show the cross section of a vortex ring moving to the right, down the x axis. This is essentially Figure (12-15) of the last chapter, which shows how the velocity field of the top half of the ring pushes the bottom half forward, while the velocity field of the bottom half pushes the top half forward. Because the velocity decreases as we go away from the core, the bigger the ring becomes, the farther the halves are apart, the slower the ring moves.

In Figure (11), we show the same vortex ring, but now we are assuming that there is a charged fluid in the core, and an external  $\hat{x}$  directed electric field is pushing on this charged fluid. It looks like we are attempting to accelerate the ring by pushing on it in the direction of its motion.

To see what this force does, we go back to Figure (8) and see that the  $\hat{x}$  directed force  $\vec{g}$  acting on the fluid in a + $\hat{z}$  oriented core causes the core to move up in the + $\hat{y}$  direction. At the bottom of the ring where the vorticity points in the opposite direction the same  $\hat{x}$ directed force causes the core to move down (see Exercise 2). Overall the force  $\vec{g}$  is causing the entire ring to grow in size, which results in the ring moving more slowly.

Thus we have the peculiar phenomenon that when we push on a ring in the direction the ring is moving, we make the ring bigger and slow it down. In Exercise (3), you show that if you push opposite to the direction of motion of the ring, you make the ring smaller and faster.



Cross section of a vortex ring. Each side of the ring moves the other side forward. The smaller the ring, the greater the velocity field, and the faster the ring moves.

#### **Exercise 3**

Using Equation (78), show that when you push opposite to the direction of motion of the ring you speed it up.

# **Conservation of Energy**

At first sight you might think you have a problem with the law of conservation of energy when it comes to the behavior of vortex rings. When we push on an object in the direction that it is moving, we are doing positive work on the object, and expect that, in the absence of friction, the energy of the object would increase. But for a vortex ring, when we push in the direction of the ring's motion the ring slows down. Does the ring loose energy as a result?

No. Unlike baseballs and other objects we are familiar with, a vortex ring's kinetic energy increases when it slows down. That is because its diameter increases and thus there is more length of vortex line. The kinetic energy of the ring is the kinetic energy  $1/2 \text{ mv}^2$  of the fluid particles whose motion is caused by the ring. The larger the ring, the more fluid involved in the vortex motion, and the more kinetic energy associated with the ring. Thus pushing on a ring in the direction of motion increases its energy, as it should.





An  $\vec{x}$  directed force acting on a ring moving in the  $\vec{x}$  direction causes the ring to expand.



**Figure 12** *Pushing opposite to the direction of motion of the ring.* 

# Measurement of the Quantized Circulation $\kappa = h/m_{\rm He}$

We have mentioned that Rayfield and Reif could control and measure the behavior of their charged vortex rings by sending pulses of the rings between grids in the superfluid. By timing the pulse, they could measure the speed of the rings. By applying a voltage difference to the grids, they could change the energy of the rings. A voltage difference V<sub>voltage</sub> would cause an energy change of magnitude ( $eV_{voltage}$ ) for each ring because each ring carried either one proton of charge (+e) or one electron of charge (-e). We will give a rough argument as to how these two kinds of measurements allowed Rayfield and Reif to accurately measure the quantized circulation  $\kappa = h/m_{He}$  of the ring.

We have noted that the energy of a ring is the kinetic energy  $1/2 \text{ mv}^2$  of the fluid particles. Since the velocity field of a vortex is proportional to the vortex's circulation  $\kappa$  ( $\vec{v} = \kappa/2\pi r$  for a straight vortex), the fluid kinetic energy is proportional to  $\kappa^2$ . The fluid energy in a vortex ring is also proportional to the length  $2\pi R$  of line in the ring. As a result the fluid kinetic energy is proportional to  $\kappa^2 R_{\text{ring}}$ 

 $E_{\text{ring}} \propto \kappa^2 R_{\text{ring}} \qquad \begin{array}{c} \text{kinetic energy} \\ \text{of } a \\ \text{vortex ring} \end{array} \tag{83}$ 

#### Exercise 4

Show that  $\rho\,\kappa^2 {\sf R}_{\,\text{ring}}$  has the dimensions of kinetic energy.

We have seen that the velocity of a pair of oppositely oriented vortices is given by the formula

$$V_{2D ring} = \frac{\kappa}{4\pi R_{ring}}$$
(12-40)

and have noted that the speed of a circular ring is roughly the same but more complex. In any case it is proportional to  $\kappa/R_{ring}$ 

$$V_{\text{ring}} \propto \frac{\kappa}{R_{\text{ring}}} \qquad speed of \\ vortex ring \qquad (84)$$

Neither Equation (83) or (84), or an accurate calculation of these quantities, can be used to measure the circulation  $\kappa$  of the ring because you cannot see the rings to measure their radius R<sub>ring</sub>. But in the

product of the two terms, the unmeasurable term  $R_{ring}$  cancels and we are left with the formula

$$E_{ring} \times V_{ring} \propto \kappa^3$$
 (85)

Equation (85) suggests that an experimental measurement of  $E_{ring} \times V_{ring}$  will give an experimental value of  $\kappa^3$ . A careful (and messy) calculation shows that both  $E_{ring}$  and  $V_{ring}$  have factors of the logarithm of the ring radius  $R_{ring}$  divided by the core diameter (a). As a result there are factors of ln( $R_{ring}/a$ ) in a more accurate formula for the product  $E_{ring} \times V_{ring}$ . However this logarithm is quite insensitive to the actual value of  $R_{ring}/a$  (increase the ring radius by 1000 and the logarithm ln( $R_{ring}/a$ ) increases only by an additional amount of 6.9). By making a number of measurements of  $E_{ring} \times V_{ring}$ , Rayfield and Reif were not only able to determine  $\kappa$ , but also the core diameter (a). That is when they found that the core diameter was roughly the diameter of a helium atom.

## **The Magnus Equation**

In Figure (8) repeated here we show a  $\hat{z}$  directed vortex line, subjected to an  $\hat{x}$  directed force, moving in the  $\hat{y}$  direction. This motion labeled  $\vec{V}_{rel}$  is the motion of the line relative to the fluid particles due to the non potential force  $\vec{g}$ . For the special case of a straight vortex, it is fairly easy to calculate the magnitude of this relative velocity  $\vec{V}_{rel}$ . The result we will call the *Magnus equation*, named after a person who first studied sideways motion due to vortex effects.



**Figure 8 (repeated)** The y directed motion of a z oriented vortex line subject to an x directed force.

For this calculation, assume that we have a core of diameter D, with a uniform  $\hat{z}$  directed vorticity  $\vec{\omega}$  and an  $\hat{x}$  directed force inside, as shown in Figure (13). We have drawn two paths C'<sub>1</sub>(t) and C'<sub>2</sub>(t) attached to the fluid particles. The circuits nearly touch each other so that half of the flux of  $\vec{\omega}$  goes through C'<sub>1</sub> and half through C'<sub>2</sub> at the time (t).

A little time  $\delta t$  later, the core has moved upward a distance  $\delta y$  relative to the fluid particles as shown in Figure (13b). To keep the calculation simple, we will assume that the force  $\vec{g}$  is strong enough to move the core up a reasonable distance  $\delta y$  before the fluid has moved the circuits C'<sub>1</sub> and C'<sub>2</sub> noticeably. (The more accurate calculation in Appendix 2 does not make this assumption, but gets the same answer.)

Because the vorticity is moving up relative to the fluid particles, and thus up relative to the circuits  $C'_1$  and  $C'_2$ , by the time  $(t+\delta t)$  we have an additional band of flux of area  $(D \,\delta y)$  through circuit  $C'_1$ . Thus the increase  $\delta \Phi_1$  of flux in circuit  $C'_1$ , as we go from (t) to  $(t+\delta t)$ , is

$$\delta \Phi_1 = \omega(D \,\delta y) \tag{86}$$

Applying our vortex dynamics Equation (78) to the upper circuit  $C'_1$ , we have

$$\frac{\delta \Phi_1}{\delta t} = \oint_{C'_1} \vec{g} \cdot d\vec{\ell} \qquad {rate of increase} \\ of flux of \vec{\omega} \\ through C'_1$$
(87)

Looking at Figure (13a) we see that the only contribution we get to  $\oint \vec{g} \cdot d\vec{\ell}$  around C'<sub>1</sub> is through the center of the core, where  $\vec{g}$  acts for a distance D, giving



#### Figure 13

As the core moves up relative to the fluid particles, and thus up relative to the paths  $C'_1$  and  $C'_2$  attached to the fluid particles, we get at time  $(t + \delta t)$  an additional band of flux of area  $(D \delta y)$  in circuit  $C'_1$ .

$$\oint_{C_1'} \vec{g} \cdot d\vec{\ell} = gD$$
(88)

Thus

$$\frac{\delta \Phi_1}{\delta t} = gD \quad ; \quad \delta \Phi_1 = gD\delta t \tag{89}$$

Equating the values of  $\delta \Phi_1$  from Equations (86) and (89) gives

$$\delta \Phi_1 = \omega D \, \delta y = g D \, \delta t \tag{90}$$

The D's cancel, and we are left with

$$g = \omega \frac{\delta y}{\delta t} = \omega V_{rel}$$
(91)

where  $V_{rel}$  is the relative velocity of the vortex core and the fluid particles.

Equation (91) can be put in a more useful form if we multiply both sides by  $\rho$ , converting the force g per unit mass to  $\rho g = f$ , the force per unit volume. Then integrate f over the area of the core, giving us the *force per unit length* acting on the core. We get, using Equation (91)  $g = \omega V_{rel}$ ,

$$F_{e} = \int_{\substack{\text{area of} \\ \text{core}}} \rho g dA = \rho \int_{\substack{\text{area of} \\ \text{core}}} (\omega V_{rel}) dA$$
$$= \rho V_{rel} \int_{\substack{\text{area of} \\ \text{core}}} \omega dA$$
(92)

But the integral of  $\omega$  over the area of the core is  $\kappa$ , the total circulation of the core. Thus Equation (92) becomes

$$F_e = \rho \kappa V_{rel} \tag{93}$$

The final step is to turn Equation (93) into a vector equation. We let the vector  $\vec{\kappa} = \hat{z}\kappa$  point in the direction of the vorticity  $\vec{\omega}$ . The force  $\vec{F}_e$  points in the  $\hat{x}$  direction and  $\vec{V}_{rel}$  is  $\hat{y}$  directed. Using the right hand rule, we see that the cross product  $\vec{V}_{rel} \times \vec{\kappa}$  points in the  $\hat{x}$  direction like  $\vec{F}_e$ . Thus we have the vector equation

$$\vec{F}_{e} = \rho \vec{V}_{rel} \times \vec{\kappa} \qquad \begin{array}{c} Magnus\\ equation \end{array} \tag{94}$$

which is a remarkably simple result for what looked like a complex situation.

In Appendix 2 to this chapter, we derive an equation for the effect of non potential forces on curved fluid core vortices. The result looks exactly like Equation (94), but it tells us how to define  $\vec{V}_{rel}$  when we have a curved vortex.

$$\vec{F}_{e} = \rho \vec{V}_{rel} \times \vec{\kappa}$$
 (94) repeated

When the exact formula is applied to a straight vortex in a two dimensional flow, the terms in Equation (94) have the following meaning. If  $\hat{z}$  is the direction perpendicular to the flow, then  $\vec{F}_e$  is the x-y component of the total force per unit length acting on the fluid in the core region. The component  $(F_e)_z$  parallel to the vortex has no effect. The circulation  $\vec{\kappa}$  is the total flux of  $\vec{\omega}$  in the core, and is  $\hat{z}$  oriented.

The relative velocity  $\vec{V}_{rel}$  is given by the formula

$$\vec{\mathbf{V}}_{\text{rel}} = \vec{\mathbf{V}}_{\text{vortex}} - \vec{\mathbf{V}}_{\text{fluid}}$$
 (95)

where the vortex velocity  $\vec{V}_{vortex}$  is the velocity of the center of mass of the vorticity  $\omega_z$ , and the fluid velocity  $\vec{V}_{fluid}$  is the weighted average of the fluid velocity  $\vec{v}$  in the core region, given by the integral



Figure 14

Relative directions of  $\vec{\omega}$ ,  $\vec{F}_e$ , and  $\vec{V}_{rel}$ .

With these definitions, Equation (94) is an exact equation for a straight fluid core vortex. The result is independent of the shape of the core or the force density  $\vec{g}$ , as long as both are confined to a localized region.

The derivation of the exact Magnus equation, which we do in Appendix 2, is obtained by going back to Equation (55) and rewriting that equation as a continuity equation for the flow of vorticity. In some ways the continuity equation is simpler to derive and use than the Helmholtz theorem approach. But the continuity equation involves the quantity  $\varepsilon_{ijk}$  which we introduce and use in Appendix 1 to derive various vector identities. Thus it seemed appropriate to delay a discussion of the continuity equation until after the reader has studied Appendix 1.

(The beginning of Appendix 2 gives a complete physical explanation of the continuity equation approach with virtually no mathematics and can be read at any time.)

# **IMPULSE OF A VORTEX RING**

Although we have discussed the Magnus equation  $\vec{F} = \rho \vec{V}_{rel} \times \vec{\kappa}$  as applied to a straight vortex, the same ideas can be used for a curved vortex as long as the radius of curvature of the vortex is large compared to the core radius. When we apply the Magnus equation to a vortex ring, we get a simple formula relating the total force on the ring to the rate of change of the area of the ring. Introducing the concept of the *impulse of a vortex ring*, we can write this formula so that looks a lot like Newton's law for vortex rings.

In Figure (15) we again show the cross section of a vortex ring, now showing the force  $\vec{F}_e$  per unit length acting on each section of the core, and the relative velocity  $\vec{V}_{rel}$  causing the ring to expand. For simplicity let  $\vec{F}_e$  be in the direction of the motion of the ring, so that the Magnus equation implies

$$F_{\rm e} = \rho \kappa V_{\rm rel} \tag{94a}$$

The velocity  $V_{rel}$  is just the rate  $dR_{ring}/dt$  that the ring radius is increasing. Thus Equation (94a) becomes

$$F_{\rm e} = \rho \kappa \frac{\mathrm{d}R_{\rm ring}}{\mathrm{d}t} \tag{97}$$

The  $F_e$  in Equation (97) is the force per unit length of the ring. The total length of the ring is its circumference  $2\pi R_{ring}$ , thus the total force  $F_{total}$  is  $2\pi R_{ring}F_e$ , giving

$$F_{\text{total}} = 2\pi R_{\text{ring}} F_{\text{e}} = 2\pi\rho\kappa \left[ R_{\text{ring}} \frac{dR_{\text{ring}}}{dt} \right] \quad (98)$$

However

$$R\frac{dR}{dt} = \frac{1}{2}\frac{d}{dt}(R^2)$$
(99)



Thus Equation (98) can be written in the form

$$F_{\text{total}} = \rho \kappa \frac{d}{dt} (\pi R_{\text{ring}}^2)$$
(100)

But  $\pi R_{ring}^2$  is just the area  $A_{ring}$  of the ring, thus we get

$$F_{\text{total}} = \frac{d}{dt} (\rho \kappa A_{\text{ring}})$$
(101)

Let us define the vector  $\vec{A}_{ring}$  as a vector of magnitude  $\pi R_{ring}^2$ , pointing in the direction of the motion of the ring. Then since the total force  $\vec{F}_{total}$  also points in the same direction, we can write Equation (101) as the vector equation

$$\vec{F}_{total} = \frac{d}{dt} (\rho \kappa \vec{A}_{ring})$$
 (102)

Of course we have derived Equation (102) only for the special case that  $\vec{F}_{total}$  points in the direction the ring is moving. It becomes an interesting exercise with the vector form of the Magnus equation to show that Equation (102) applies for any direction of  $\vec{F}_{total}$ .

Equation (102) seems to look a lot like Newton's second law relating the total force  $\vec{F}$  acting on a particle to the particle's momentum  $\vec{p}$ 

$$\vec{F} = \frac{d\vec{p}}{dt}$$
 Newton's second law

Equation (102) suggests that the quantity  $\rho \kappa \vec{A}_{ring}$ plays a role for vortex rings similar to the role of momentum for particles. As a result it has become traditional to give  $\rho \kappa \vec{A}_{ring}$  a special name, the *impulse*  $\vec{I}$  of the ring

$$\vec{I} \equiv \rho \kappa \vec{A}_{ring}$$
 impulse of a (103)

With Equation (103) the formula for  $\vec{F}_{total}$  becomes

$$\vec{F}_{total} = \frac{d\vec{I}}{dt}$$
 *impulse* (104)

A common error one can make is to associate the impulse  $\vec{I}$  of a vortex ring with an actual fluid momentum. Suppose, for example, you have a vortex ring in a sealed container. If you integrate  $\rho \vec{v}$  for that ring over the entire fluid, the answer is *zero*! In other words *vortex rings do not carry linear momentum*. The impulse  $\vec{I}$  is a separate quantity with its own special properties. One important property is that it makes it easy to predict the behavior of a ring subject to external forces. But it is not the momentum of the ring.

# THE AIRPLANE WING

In the fluid dynamics Chapter 23 of the Physics text, we used Bernoulli's equation to provide a qualitative view of why airplanes fly and sailboats can sail into the wind. In this section we will first look at the flow pattern of the fluid past an airplane wing, and see that for there to be lift, there has to be a net circulation of the fluid around the ring. This means that there is a vortex surrounding the wing. We then use the Magnus equation (95) to obtain a formula relating the weight of the airplane to the forward speed of the airplane and the circulation  $\kappa$  of the vortex about the wing.

Figure (16) is a sketch of the streamlines we might expect for the flow of a fluid past an airplane wing. Our Bernoulli equation argument was that because the fluid was flowing faster over the top of the wing (where the streamlines are closer together) and slower under the wing, the pressure must be higher under the wing than on top so that the sum of the terms  $(p + \rho v^2/2)$  be constant. (The  $\rho$ gy term is too small to worry about for a fluid like air.) This higher pressure below suggests that the fluid is exerting a lift force on the wing. In Figure (16) we have drawn a circuit C' around the wing. When we calculate the integral  $\oint \vec{v} \cdot d\vec{\ell}$  around this circuit, we get a big positive contribution from the high speed fluid at the top, and a smaller negative contribution from the slow fluid at the bottom. Thus there is a net positive circulation  $\vec{\kappa}$  surrounding the wing. In Figure (16), the circulation  $\vec{\kappa}$  points in the  $+\hat{z}$ direction. If there were no net circulation, if the fluid had the same speeds above and below the wing, there would be no lift.



Here is where we will adopt a rather unconventional view in order to directly apply the Magnus equation (94) to the airplane wing. We will picture the wing as being made of frozen fluid of the same density as the air flowing over it. This way we can think of the wing itself as part of the fluid, giving us a constant density, fluid core vortex to which we can apply Equation (95). Because the Magnus equation involves only the total circulation  $\vec{k}$  and not the details of the structure of the core, it makes no difference that our core now consists of a vortex sheet around the surface of the wing rather than the solid-body like rotation we assumed in our other vortex cores.

The purpose of the wing is to support the weight  $m\vec{g}$  of the airplane. If we divide  $m\vec{g}$  by the total length L of the wings, we get the downward,  $-\hat{y}$  directed force  $\vec{F}_g$  per unit length acting on the wings, and thus on the core of the wing vortex.

Here is the unconventional part of the argument. If you exert a downward,  $-\hat{y}$  directed force on a  $\hat{z}$  oriented vortex, you will get an  $\hat{x}$  directed relative velocity of the core as shown in Figure (17). (Figure (17) is just Figure (8) rotated 90°.) Comparing Figures (16) and (17), we can say that the *downward gravitational force* on the wing, i.e., on the core of the vortex around the wing, is *causing the wing vortex to move forward relative to the fluid* through which the airplane is flying. The Magnus equation, with  $\vec{F}_e = \vec{F}_g$  is

$$\vec{F}_{g} = \rho \vec{V}_{rel} \times \vec{\kappa} \tag{104}$$

This gives us an explicit formula relating the downward gravitational force  $\vec{F}_g$  per unit length, the circulation  $\vec{\kappa}$  of the wing vortex, and the forward speed  $\vec{V}_{rel}$  of the airplane.



**Figure 17** Motion of a vortex subject to a localized force  $\vec{g}$ .

The first thing this equation tells you is that there must be a vortex around the wing of an airplane for the airplane to fly. In addition, the vortex cannot stop at the end of the wing because vortex lines, being solenoidal  $(\vec{\nabla} \cdot \vec{\omega} = 0)$ , cannot stop in the fluid. Instead the vortices trail back behind the airplane and are sometimes very visible during takeoff on a misty morning.

Equation (94) also tells us that for a given speed  $\vec{V}_{rel}$ , the heavier the airplane, i.e., the greater  $\vec{F}_g$  is, the greater the circulation  $\vec{\kappa}$  has to be. To lift the airplane, the circulation has to be particularly strong during takeoff where the forward velocity  $\vec{V}_{rel}$  of the airplane is small. As a result the massive jumbo jets have strong wing tip vortices trailing after them, strong enough to flip small airplanes taking off behind them. Pilots of small aircraft are warned to stay clear of the jumbo jets.

We have just presented the rather different picture that the forward motion of an airplane is caused by the gravitational force acting down on the core of the wing vortex. When this point of view was presented in a science journal article, a reviewer replied that it was the airplane motors which pulled the airplane forward. Our response to that was—what about a glider that flies without motors? The main role of the motors in level flight is to overcome the viscous drag on the wings and fuselage.

Although it works well, our picture is still unconventional. When we used the Bernoulli argument in Chapter 23 of the Physics text, we were using the conventional picture that the fluid is exerting a lift force on the wing. The conventional derivation of the lift force involves calculating the momentum transfer between the fluid and the solid object. This is a somewhat messy calculation involving integration of pressure forces over the surface of the object. When you finish, you find that the lift force is proportional to the total circulation  $\vec{k}$  about the wing and the velocity  $\vec{V}_{rel}$  of the wing relative to the fluid through which it is moving. Such a lift force on a moving vortex is called the *Magnus Force*.

#### The Magnus Lift Force

We are in a position to write down the formula for the lift force on an airplane wing without doing any pressure force integrations. Start with Equation (104)

$$\vec{F}_{g} = \rho \vec{V}_{rel} \times \vec{\kappa}$$
 (104) repeated

which relates the gravitational force  $\vec{F}_g$  per unit length to the circulation  $\vec{\kappa}$  and the relative velocity  $\vec{V}_{rel}$  of the vortex. If the plane is in level flight, then the downward gravitational force  $\vec{F}_g$  must be exactly balanced by the upward lift force  $\vec{F}_{lift}$  for the plane not to rise or fall. Thus we have

$$\vec{F}_{\text{lift}} = -\vec{F}_{\text{g}} \tag{105}$$

which gives us

$$\vec{F}_{\text{lift}} = -\rho \vec{V}_{\text{rel}} \times \vec{\kappa} \tag{106}$$

In addition to airplane wings, spinning objects generally have a vortex around them. If the object is moving through the fluid at a velocity  $\vec{V}_{rel}$ , it will experience a sideways lift force given by Equation (106). This sideways lift force on a spinning object is called the Magnus force  $\vec{F}_{magnus}$  after G. Magnus who studied the sideways motion of spinning objects in 1852\*. The Magnus lift force formula found in textbooks is

$$\vec{F}_{magnus} = -\rho \vec{V}_{rel} \times \vec{\kappa} \qquad \begin{array}{c} Magnus \\ lift force \\ formula \end{array} (107)$$

\* "On the deviation of projectiles; and on a remarkable phenomenon of rotating bodies." G. Magnus, Memoirs of the Royal Academy, Berlin(1852). English translation in Scientific Memoirs, London(1853)., p.210. Edited by John Tyndall and William Francis.

# **The Magnus Force and Fluid Vortices**

The extended Helmholtz theorem, Equation (78) and its application to the motion of vortex lines through a fluid, was developed in the 1960s to help understand vortex behavior in the Rayfield-Reif experiment. Before that, and still in most textbooks, the motion of vortices through a fluid is explained in the following way.

The Magnus force formula  $\vec{F}_{magnus} = -\rho \vec{V}_{rel} \times \vec{\kappa}$  tells us the lift force on a solid object moving through a fluid at a velocity  $\vec{V}_{rel}$ , when there is a circulation  $\kappa$  about the object.

If one has a fluid core vortex moving relative to the fluid, one says that there must be a lift force  $\vec{F}_{magnus} = -\rho \vec{V}_{rel} \times \vec{\kappa}$  acting on that vortex. But there is no extra mass associated with a fluid core vortex, so one must treat the vortex as a massless object, with the result that the net force on the vortex must be zero. That means that there must be an external force  $\vec{F}_{external}$  acting on the vortex to cancel the Magnus lift force. That is, one must have

$$\vec{F}_{external} + \vec{F}_{magnus} = 0$$
 (108)

Using the Magnus formula (107) in (108) gives

$$\vec{F}_{external} = -\vec{F}_{magnus} = \rho \vec{V}_{rel} \times \vec{\kappa}$$
 (109)

This is just our Equation (95) relating the relative motion of a vortex to the localized, non potential force on the core of the vortex.

What we have shown, by deriving Equation (109) directly from the Navier-Stokes equation, which itself came from Newton's second law, is that we can describe vortex motion without any reference what-soever to a Magnus lift force. The Magnus force is a *pseudo force*, which like the centrifugal force, may be very useful for calculation, but which has no place in a basic description of the motion of the fluid itself.

# Appendix for Chapter 13 Part 1

# Component Notation and the Functions $\delta_{ij}$ and $\epsilon_{ijk}$

In our derivation of the Navier-Stokes equation we ran into the term  $\nabla_i(\vec{p}v_i)$  which we could not handle very well with vector notation like  $\vec{\nabla} \cdot \vec{v}$  or  $\vec{\nabla} \phi$ . To handle this term we resorted to component notation  $\nabla_i$  and  $v_i$ , and introduced the Einstein summation convention. Here we will briefly review the summation convention, and then discuss two quantities  $\delta_{ij}$ and  $\epsilon_{ijk}$  that play basic roles when we work with dot and cross products in component notation. These quantities also become extremely useful when we are working out vector identities, like the relationship

$$(\vec{v} \cdot \vec{\nabla}) \vec{v} = \vec{\nabla} \left( \frac{v^2}{2} \right) - \vec{v} \times (\vec{\nabla} \times \vec{v})$$
 (13-33)

which we used to get the  $v^2/2$  term in Bernoulli's equation.

#### THE SUMMATION CONVENTION

In Equation (12) of this chapter we wrote the dot product of two vectors  $\vec{a} \cdot \vec{b}$  in the following three forms

$$\vec{a} \cdot \vec{b} = a_x b_x + a_y b_y + a_z b_z$$
$$= \sum_{i=x,y,z} a_i b_i$$
(13-12)
$$= a_i b_i$$

With the summation convention, when we have repeated indices like  $a_ib_i$ , it is understood that we are to sum over all values of the repeated index i. We gave as an example

$$a_i b_j c_i = a_x b_j c_x + a_y b_j c_y + a_z b_j c_z$$

where we summed over the repeated index i, but the single index j was not summed. In mixed index-vector notation,  $a_i b_j c_j$  could be written

$$(a_i b c_i)_j = a_i b_j c_i \tag{1}$$

# THE DOT PRODUCT AND $\delta_{ij}$

We will see that the quantity  $\delta_{ij},$  defined by the simple relationship

$$\delta_{ij} = 1 \quad \text{if } i = j$$
  
= 0 \quad if \quad i \neq j (2)

is closelly related to the dot product in component notation. Consider the term

$$\delta_{ij}a_ib_j$$
 (3)

Here both indices i and j are repeated, so that we have to sum over both to get

$$\begin{split} \delta_{ij}a_{i}b_{j} &= \delta_{xx}a_{x}b_{x} + \delta_{xy}a_{x}b_{y} + \delta_{xz}a_{x}b_{z} \\ &+ \delta_{yx}a_{y}b_{x} + \delta_{yy}a_{y}b_{y} + \delta_{yz}a_{y}b_{z} \quad (4) \\ &+ \delta_{zx}a_{z}b_{x} + \delta_{zy}a_{z}b_{y} + \delta_{zz}a_{z}b_{z} \end{split}$$

In Equation (4), the only non zero  $\delta_{ij}$  terms are  $\delta_{xx}$ ,  $\delta_{yy}$  and  $\delta_{zz}$ , leaving

$$\delta_{ij}a_ib_j = \delta_{xx}a_xb_x + \delta_{yy}a_yb_y + \delta_{zz}a_zb_z \quad (5)$$

Since  $\delta_{xx} = \delta_{yy} = \delta_{zz} = 1$ , we get

$$\delta_{ij}a_ia_j = a_xb_x + a_yb_y + a_zb_z = \vec{a}\cdot\vec{b}$$
(6)

In component notation this can be written

$$\delta_{ij}a_ib_j = a_jb_j = \vec{a}\cdot\vec{b}$$
(7)

You can see that the function  $\delta_{ij}$  turns the product of two vectors  $a_i$  and  $b_j$  into a dot product.

Another way of handling  $\delta_{ij}a_ib_j$  is to first work out the effect of  $\delta_{ij}$  acting on  $a_i$ . Setting the index j to x we have

$$\delta_{ix}a_i\ =\ \delta_{xx}a_x+\delta_{xy}a_y+\delta_{xz}a_z\ =\ a_x$$

Similarly we get

$$\delta_{iy}a_i = a_y$$
  
 $\delta_{iz}a_i = a_z$ 

Thus for any value of j,  $\delta_{ij}a_i$  is equal to  $a_j$ 

$$\delta_{ij}a_i = a_j \tag{8}$$

Then when we want to evaluate the product  $\delta_{ij}a_ib_j$  we can write

$$(\delta_{ij}a_i)b_j = (a_j)b_j = \vec{a}\cdot\vec{b}$$
(9)

# THE CROSS PRODUCT AND $\epsilon_{ijk}$

We just saw that  $\delta_{ij}$  turned the product of two vectors  $a_i$  and  $b_j$  into a dot product  $\vec{a} \cdot \vec{b}$ . We will now see that a slightly more complex function  $\epsilon_{ijk}$  turns the product of two vectors  $a_i b_k$  into a cross product  $\vec{a} \times \vec{b}$ 

The cross product  $\vec{a} \times \vec{b}$  of two vectors is given by

$$(\vec{a} \times \vec{b})_{x} = a_{y}b_{z} - a_{z}b_{y}$$

$$(\vec{a} \times \vec{b})_{y} = a_{z}b_{x} - a_{x}b_{z}$$

$$(10)$$

$$(\vec{a} \times \vec{b})_{z} = a_{x}b_{y} - a_{y}b_{x}$$

We will see that this can all be written as the one equation

$$(\vec{a} \times \vec{b})_i = \varepsilon_{ijk} a_j b_k \tag{11}$$

where the function  $\varepsilon_{iik}$  has the values

$$\begin{split} \epsilon_{ijk} &= 0 \quad \text{if any two indices are equal} \\ \epsilon_{xyz} &= 1 \\ \epsilon_{xzy} &= -1 \\ \epsilon_{zxy} &= +1 \end{split} \tag{12}$$

What we are indicating by the dots is that if you permute (interchange) any two neighboring indices, you change the sign.

For example, what is the sign of  $\varepsilon_{zyx}$ ? To find out we do the following permutations starting with  $\varepsilon_{xyz} = +1$ 

$$\varepsilon_{xyz} = +1$$

$$\varepsilon_{xzy} = -1$$

$$\varepsilon_{zxy} = +1$$

$$\varepsilon_{zyx} = -1$$
(13)

It does not matter how you do the permutation you always come out with the same answer. For example

$$\epsilon_{xyz} = +1$$

$$\epsilon_{yxz} = -1$$

$$\epsilon_{yzx} = +1$$

$$\epsilon_{zyx} = -1$$
(13a)

Because of this permutation property,  $\varepsilon_{ijk}$  is often called the *permutation tensor*. (A *tensor* is a vector like object with more than one index.)

Now we have to check that Equation (11), using  $\varepsilon_{ijk}$  for the cross product, gives the correct result. Using the summation convention and crossing out terms like  $\varepsilon_{xxk}$  which are zero, we have

$$(\vec{a} \times \vec{b})_{x} = \varepsilon_{xjk} a_{j} b_{k}$$
  
=  $\varepsilon_{xxk} a_{x} b_{k} + \varepsilon_{xyk} a_{y} b_{k} + \varepsilon_{xzk} a_{z} b_{k}$   
=  $\varepsilon_{xyx} a_{y} b_{x} + \varepsilon_{xyy} a_{y} b_{y} + \varepsilon_{xyz} a_{y} b_{z}$   
+  $\varepsilon_{xzx} a_{z} b_{x} + \varepsilon_{xzy} a_{z} b_{y} + \varepsilon_{xzz} a_{z} b_{z}$ 

$$(\vec{a} \times \vec{b})_{x} = \varepsilon_{xyz} a_{y} b_{z} + \varepsilon_{xzy} a_{z} b_{y}$$
(14)

With  $\varepsilon_{xyz} = +1$ ,  $\varepsilon_{xzy} = -1$  (one permutation), we get

$$(\vec{a} \times \vec{b})_{x} = a_{y}b_{z} - a_{z}b_{y}$$
(15)

Which is the correct answer.

Exercise 1 Check that

$$(\vec{a} \times \vec{b})_y = e_{yjk} a_j b_k$$

As an example of the use of the  $\epsilon_{ijk}$  , let us prove the vector identity

$$\vec{A} \cdot (\vec{B} \times \vec{C}) = (\vec{A} \times \vec{B}) \cdot \vec{C}$$
(13-68)

which we used in the derivation of the Helmholtz theorem. We have

$$\vec{A} \cdot (\vec{B} \times \vec{C}) = A_i (\vec{B} \times \vec{C})_i$$
$$= A_i \epsilon_{ijk} B_j C_k$$
$$\vec{A} \cdot (\vec{B} \times \vec{C}) = \epsilon_{ijk} A_i B_j C_k$$
(16)

$$(\vec{A} \times \vec{B}) \cdot \vec{C} = (\vec{A} \times \vec{B})_{i}C_{i}$$
$$= \epsilon_{ijk}A_{j}B_{k}C_{i}$$
$$(\vec{A} \times \vec{B}) \cdot \vec{C} = \epsilon_{ijk}A_{j}B_{k}C_{i}$$
(17)

To show that Equation (17) is equivalent to (16), we will first rename the indices in Equation (16). We will do this in two steps to avoid any possible errors. Changing  $i \rightarrow r$ ,  $j \rightarrow s$ ,  $k \rightarrow t$  in Equation (16) gives

$$\varepsilon_{ijk}A_iB_jC_k = \varepsilon_{rst}A_rB_sC_t \tag{16a}$$

We can do this because it does not matter what letter we use for a repeated index. Now we wish to rename the indices again so that the vector components in Equation (16) match those in (17). If we substitute  $r \rightarrow j$ ,  $s \rightarrow k$ ,  $t \rightarrow i$ , Equation (16a) becomes

$$\varepsilon_{\rm rst} A_{\rm r} B_{\rm s} C_{\rm t} = \varepsilon_{\rm jki} A_{\rm j} B_{\rm k} C_{\rm i} \tag{16b}$$

which when combined with (17a) gives

$$\varepsilon_{ijk} A_i B_j C_k = \varepsilon_{jki} A_j B_k C_i \tag{16c}$$

With some practice, you will not bother going through steps (16a) and write (16b) directly.

We now have

$$\vec{A} \cdot (\vec{B} \times \vec{C}) = \varepsilon_{jki} A_j B_k C_i$$
(16d)

$$(\vec{A} \times \vec{B}) \cdot \vec{C} = \epsilon_{ijk} A_j B_k C_i$$
 (17) repeated

The vector components now match, and what we now have to do is see how  $\varepsilon_{jki}$  compares with  $\varepsilon_{ijk}$ . We will start with  $\varepsilon_{ijk}$  and see how many permutations it takes to get to  $\varepsilon_{jki}$ . We have

$$\begin{split} \epsilon_{jik} &= -\epsilon_{ijk} \\ \epsilon_{jki} &= -\epsilon_{jik} = -(-\epsilon_{ijk}) \end{split}$$

Two permutations are required, we have  $\varepsilon_{jki} = \varepsilon_{ijk}$ , and thus the terms in (16) and (17) are equal, which proves the identity.

While these steps may have looked a bit complex the first time through, with some practice they are much easier, faster, and more accurate than writing out all the x, y, and z components of the cross products.

#### **Handling Multiple Cross Products**

To work out vector identities involving more than one cross product, there is a special identity that is worth memorizing. It is

$$\varepsilon_{ijk}\varepsilon_{klm} = \delta_{il}\delta_{jm} - \delta_{im}\delta_{jl}$$
(18)

First of all, note that Equation (18) has the correct symmetry. It must change sign on the right if you permute (interchange) i and j or l and m, because that is what  $\varepsilon_{ijk}$  and  $\varepsilon_{klm}$  do on the left side. This combination of  $\delta$  functions has that property.

Before we try to prove Equation (18), we will give an example of how useful it is. Consider the rather messy set of cross products  $\vec{a} \times (\vec{b} \times \vec{c})$ . Using the  $\varepsilon_{iik}$  notation for cross products, we have

$$\left[ \vec{a} \times (\vec{b} \times \vec{c}) \right]_{i} = \epsilon_{ijk} a_{j} (\vec{b} \times \vec{c})_{k}$$
$$= \epsilon_{ijk} a_{j} \epsilon_{klm} b_{l} c_{m} \qquad (19)$$
$$= (\epsilon_{ijk} \epsilon_{klm}) a_{j} b_{l} c_{m}$$

Using Equation (18) we get

$$\left[\vec{a} \times (\vec{b} \times \vec{c})\right]_{i} = (\delta_{il}\delta_{jm} - \delta_{im}\delta_{jl})a_{j}b_{l}c_{m} (20)$$

We will get some practice with the use of the  $\delta$  functions  $\delta_{ii}$ . We have for example

$$\delta_{il}b_1 = b_i \quad ; \quad \delta_{jm}c_m = c_j \tag{21}$$

So that

$$\delta_{il}\delta_{jm}a_jb_lc_m = a_jb_ic_j$$

and

$$-\delta_{im}\delta_{jl}a_{j}b_{l}c_{m} = -a_{j}b_{j}c_{i}$$

We get the result

$$\left[\vec{a} \times (\vec{b} \times \vec{c})\right]_{i} = a_{j}b_{i}c_{j} - a_{j}b_{j}c_{i}$$
(22)

To apply Equation (22) to the problem we had with the Navier-Stokes equation, let

$$\vec{a} = \vec{v}$$
;  $\vec{b} = \vec{\nabla}$ ;  $\vec{c} = \vec{v}$  (23)

giving

$$\left[\vec{v} \times (\vec{\nabla} \times \vec{v})\right]_{i} = v_{j} \nabla_{i} v_{j} - v_{j} \nabla_{j} v_{i}$$
(24)

By not changing the order of the vectors in Equation (22), the equation can be used when one or more of the vectors are the gradient vector  $\vec{\nabla}$ .

To get Equation (24) into the form we want, consider

$$\frac{1}{2}\nabla_{i}v^{2} = \frac{1}{2}\nabla_{i}(v_{x}^{2} + v_{y}^{2} + v_{z}^{2})$$

$$= \frac{1}{2}(2v_{x}\nabla_{i}v_{x} + 2v_{y}\nabla_{i}v_{y} + 2v_{z}\nabla_{i}v_{z}) \quad (25)$$

$$= v_{j}\nabla_{i}v_{j}$$

Thus Equation (24) can be written

$$\left[\vec{\mathbf{v}} \times (\vec{\nabla} \times \vec{\mathbf{v}})\right]_{i} = \nabla_{i} \left(\frac{\mathbf{v}^{2}}{2}\right) + v_{j} \nabla_{j} v_{i}$$
(26)

To put this in pure vector notation, notice that Equation (26) is the (i)th component of the vector equation

$$\vec{\mathbf{v}} \times (\vec{\nabla} \times \vec{\mathbf{v}}) = \vec{\nabla} \left( \frac{\mathbf{v}^2}{2} \right) + (\vec{\mathbf{v}} \cdot \vec{\nabla}) \vec{\mathbf{v}}$$
 (27)

Equation (27) is equivalent to

$$(\vec{\mathbf{v}}\cdot\vec{\nabla})\vec{\mathbf{v}} = -\vec{\nabla}\left(\frac{\mathbf{v}^2}{2}\right) + \vec{\mathbf{v}}\times\vec{\boldsymbol{\omega}}$$
(28)

which we used to get the Bernoulli term  $-\vec{\nabla}(v^2/2)$  into the Navier-Stokes equation.

# Proof of the $\mathcal{E}$ Identity

We will use a rather brute force method to prove the identity

$$\varepsilon_{ijk}\varepsilon_{klm} = \delta_{il}\delta_{jm} - \delta_{im}\delta_{jl}$$
 (18) repeated

Let us consider the special case i = x and j = y. Then for the  $\delta$  functions we get

$$\delta_{il}\delta_{jm} - \delta_{im}\delta_{jl} = \delta_{xl}\delta_{ym} - \delta_{xm}\delta_{yl}$$
(29)

If 
$$l = x$$
,  $m = y$ , get +1 from  $\delta_{xl}\delta_{ym}$  (30a)

If 
$$l = y$$
,  $m = x$ , get  $-1$  from  $-\delta_{xm}\delta_{yl}$  (30b)

For this case i = x and j = y, the product of  $\varepsilon$ 's, becomes

$$\varepsilon_{ijk}\varepsilon_{klm} = \varepsilon_{xyk}\varepsilon_{klm} \tag{31}$$

The only non zero value for k is z giving

$$\varepsilon_{\rm xyk}\varepsilon_{\rm klm} = \varepsilon_{\rm xyz}\varepsilon_{\rm zlm} \tag{32}$$

The only value of l and m that give a non zero result are l = x, m = y and l = y, m = x. For l = x, m = y, we get  $\varepsilon_{xyz}\varepsilon_{zxy}$ . Two permutations give

$$\varepsilon_{\text{zxy}} \Rightarrow -\varepsilon_{\text{xzy}} \Rightarrow \varepsilon_{\text{xyz}} = +1$$

Thus

 $\varepsilon_{xyk}\varepsilon_{klm} = +1$  for l = x, m = y (33a)

which agrees with Equation (30a).

For the case l = y, m = x, we get

$$\varepsilon_{xyk}\varepsilon_{klm} = \varepsilon_{xyz}\varepsilon_{zyx}$$

Now

$$\varepsilon_{zyx} \Rightarrow -\varepsilon_{zxy} \Rightarrow \varepsilon_{xzy} \Rightarrow -\varepsilon_{xyz} = -1$$

thus

$$\varepsilon_{xyz}\varepsilon_{zyx} = (+1)(-1) = (-1)$$

and we have

$$\varepsilon_{xyk}\varepsilon_{klm} = -1$$
 for  $l = y, m = x$  (33b)

which agrees with Equation (30b). All other values of l and m give zero, in agreement with (30c).

You can see that Equation (18) is correct for the special case i = x, j = y. In a few more pages of essentially identical work you can, if you want, show that Equation (18) works for any values of i and j. For practice, perhaps you might try a case like i = z, j = y.

# Appendix for Chapter 13 Part 2 **Vortex Currents**

In the main part of Chapter 13, we derived the following equation that describes the behavior of vorticity in a constant density fluid.

$$\frac{\partial \vec{\omega}}{\partial t} - \vec{\nabla} \times (\vec{v} \times \vec{\omega}) = \vec{\nabla} \times \vec{g}_{np}$$
(13-55)

It turns out that there are two rather different ways to handle this equation. The one we used in the main part of the chapter was to show that

$$\frac{\delta \Phi(\vec{\omega})}{\delta t} = \int_{S'} \left[ \frac{\partial \vec{\omega}}{\partial t} - \vec{\nabla} \times (\vec{v} \times \vec{\omega}) \right] \cdot d\vec{A}$$

rate of change of the flux = of  $\vec{\omega}$  through a circuit S' that moves with the fluid

(13-74)

Thus if  $\vec{g}_{np} = 0$ , there is no change in the flux and we have Helmholtz's theorem. If there is a change in flux, we have the relative motion of the vortex lines and fluid particles that we discussed in detail.

The other approach, which we discuss in this appendix, is to turn Equation (13-55) into a continuity equation for the flow of the vorticity field  $\vec{\omega}$ . The physical idea of how we get a continuity equation is very straightforward. The mathematics requires a fairly extensive use of the tensor  $\varepsilon_{ijk}$  that we discussed in Appendix 1. That is why we have delayed the discussion of the flow of vorticity and vortex currents until this appendix.

Of the two approaches, the continuity equation approach is the more powerful. As we mentioned, it leads to an exact Magnus formula for curved fluid core vortices, a result that had not been obtained any other way. And the flow of vorticity, in the form of a vortex current tensor, appears to be playing a role in recent approaches to string theory.

# CONSERVED TWO DIMENSIONAL CURRENTS

Before we go through any mathematical steps, let us look at the physical ideas of why we should expect to find a conserved flow of vorticity, and why working with a conserved flow might give us a simple way to handle the dynamics of the vorticity field.

In Figure (1a) we have sketched several vortices of rather arbitrary shape that we imagine are moving around in a constant density fluid. When we originally drew this diagram, we were thinking of quantized vortex lines moving around in superfluid helium. But it turns out that our analysis applies to *tubes of flux* for any solenoidal field, i.e., any field like  $\vec{\omega}$  that has zero divergence. The significance of a solenoidal field is that the flux tubes cannot stop or start in the fluid. The tubes have no free ends in the fluid.



#### Figure 1

If you slice the solenoidal vortex lines with an arbitrary xy plane, the circles, representing the intersection of the lines and the plane, form the objects of a conserved two dimensional current. When a loop pulls out of the plane, as in the lower right corner, two circles of opposite orientation annihilate each other. Circles can be created or annihilated only in pairs, or come in through the edges.

In Figure (1a) we have also drawn a plane that cuts through these vortices. This is an arbitrary plane, slicing the fluid in any way we want. After drawing the plane, we then align the axis of our coordinate system so that the z axis is perpendicular to the plane. Thus we call this the  $\hat{z}$  plane.

Where a vortex tube or line comes up through the plane, we have drawn a white circle, and where it goes down through—a black circle. Because the flux tubes of a solenoidal field cannot start or stop in the fluid, the circles in the  $\hat{z}$  plane cannot appear or disappear one at a time. What can occur is that a loop may pull out of the plane as may be happening in the lower right hand corner. When this happens, a white circle and a black circle annihilate each other. If a loop enters the plane, we have the creation of a white circle-black circle pair.

If the plane extends well out beyond the region of the vortex lines, then we have a conservation law. The number of white circles minus the number of black circles is a constant.

We can go a step farther, and note that the circulation  $\kappa$  of each vortex tube is given by the formula

$$\kappa = \int_{\mathbf{S}'} \vec{\omega} \cdot d\vec{A} = \int_{\substack{\text{area of} \\ \text{intersection}}} \omega_z dA_z$$
(1)

We get the same result for  $\kappa$  no matter what  $\hat{z}$  plane we use for integrating  $\omega_z$ , as long as the  $\hat{z}$  plane cuts through the entire tube. As a result the white circles in Figure (1a) represent a net circulation  $+\kappa$  and the black circles  $-\kappa$ . If all the flux tubes of  $\vec{\omega}$  have the same circulation  $\kappa$ , then the total flux of  $\vec{\omega}$  through the plane is simply  $\kappa$  times the net number of circles, i.e., the number of white circles minus the number of black circles.

If the fluid is bounded, or the plane does not extend out beyond the region of the vortex lines, then the net number of circles can change by having vortex lines move in or out across the edges. Thus the more general conservation law is that *the rate of change of the net number of circles in a given region of the plane is equal to the rate at which circles are flowing in or out across the edges of the region*. This is a verbal statement of a continuity equation for the flow of the black and white circles.

#### **CONTINUITY EQUATION FOR VORTICITY**

To obtain the mathematical continuity equation for the flow of  $\omega_z$ , we start with the dynamic equation for vorticity, given by Equation (55) of Chapter 13 as

$$\frac{\partial \vec{\omega}}{\partial t} - \vec{\nabla} \times (\vec{v} \times \vec{\omega}) = \vec{\nabla} \times (\vec{g}_{np})$$
(13-55)

which obviously is equivalent to

$$\frac{\partial \vec{\omega}}{\partial t} = \vec{\nabla} \times (\vec{v} \times \vec{\omega} + \vec{g}_{np})$$
(13-55a)

In component notation this can be written as

$$\frac{\partial \omega_{j}}{\partial t} = \varepsilon_{jik} \nabla_{i} (\vec{v} \times \vec{\omega} + \vec{g}_{np})_{k}$$
(2)

where  $\varepsilon_{ijk}$  is the permutation tensor used in Appendix 1 to handle cross products. Using the fact that  $\varepsilon_{iik} = -\varepsilon_{ijk}$ , we get

$$\frac{\partial \omega_{j}}{\partial t} = -\nabla_{i} \left[ \epsilon_{ijk} (\vec{v} \times \vec{\omega} + \vec{g}_{np})_{k} \right]$$
(3)

Rather than try to deal with all the components in Equation (3), let us look at the z component of the equation, which becomes

$$\frac{\partial \omega_{z}}{\partial t} = -\nabla_{i} \left[ \varepsilon_{izk} (\vec{v} \times \vec{\omega} + \vec{g}_{np})_{k} \right]$$
(4)

Defining the vector  $\vec{j}(\omega_z)$  by the equation

$$j(\omega_z)_i = \epsilon_{izk} (\vec{v} \times \vec{\omega} + \vec{g}_{np})_k$$
(5)

we get the equation

$$\frac{\partial \omega_z}{\partial t} = -\vec{\nabla} \cdot \vec{j}(\omega_z)$$
(6)

which has the form of a continuity equation if we interpret  $\vec{j}(\omega_z)$  as the current vector for  $\omega_z$ .

This current vector  $\vec{j}(\omega_z)$  has the very special property that it is *two dimensional*; it has no  $\hat{z}$  component. The formula for the z component is

$$j(\omega_{z})_{z} = \varepsilon_{zzk}(\vec{v} \times \vec{\omega} + \vec{g}_{np})_{k} = 0$$
(7)

This is zero because  $\varepsilon_{zzk} = 0$ . Thus Equation (6) is the continuity equation for the two dimensional

flow of  $\omega_z$ , which is exactly what we expected from our discussion of Figure (1b).

The formula for  $\tilde{j}(\omega_z)$  still needs some simplification. The first step is to write  $\vec{v} \times \vec{\omega}$  in component notation to get

$$\varepsilon_{izk}(\vec{v} \times \vec{\omega})_k = \varepsilon_{izk} \varepsilon_{klm} v_l \omega_m$$

Next, use the relationship we proved in Appendix 1

$$\varepsilon_{ijk}\varepsilon_{klm} = \delta_{il}\delta_{jm} - \delta_{im}\delta_{jl} \qquad (A1-18)$$

to get

$$\epsilon_{izk}(\vec{v} \times \vec{\omega})_{k} = (\delta_{il}\delta_{zm} - \delta_{im}\delta_{zl})v_{l}\omega_{m}$$
  
=  $v_{i}\omega_{z} - v_{z}\omega_{i}$  (8)

The other simplification comes from noting that

$$(\hat{z} \times \vec{g}_{np})_{i} = \epsilon_{ijk} \hat{z}_{j} (g_{np})_{k} = \epsilon_{izk} (g_{np})_{k}$$
(9)

where we set  $\varepsilon_{ijk}\hat{z}_j = \varepsilon_{izk}$  because the unit vector  $\hat{z}$  has only a z component.

In Equation (5) using Equation (8) for  $\varepsilon_{izk}(\vec{v} \times \vec{\omega})_k$ and Equation (9) for  $\varepsilon_{izk}(g_{np})_k$  to get

$$\mathbf{j}(\boldsymbol{\omega}_{z})_{i} = \mathbf{v}_{i}\boldsymbol{\omega}_{z} - \mathbf{v}_{z}\boldsymbol{\omega}_{i} + (\hat{z} \times \vec{g}_{np})_{i}$$
(10)

We can simplify the interpretation by introducing the notation

$$\vec{\mathbf{v}} = (\mathbf{v}_{\mathbf{z}}, \vec{\mathbf{v}}_{\parallel}); \quad \vec{\boldsymbol{\omega}} = (\boldsymbol{\omega}_{\mathbf{z}}, \vec{\boldsymbol{\omega}}_{\parallel})$$
(11a)

where the vectors  $\vec{v}_{||}$  and  $\vec{\omega}_{||}$  are vectors representing the components of  $\vec{v}$  and  $\vec{\omega}$  parallel to the flow of  $\omega_z$ , i.e., components that lie in the  $\hat{z}$  plane.

Since the current vector  $\vec{j}(\omega_z)$  has no  $\hat{z}$  component, it has only a parallel component

$$\vec{j}(\omega_z) = \vec{j}_{\parallel}(\omega_z)$$
 (11b)

With this notation, we can let the index i be the parallel component in Equation (10), giving

$$\vec{j}(\omega_z) = \vec{v}_{\parallel}\omega_z - v_z\vec{\omega}_{\parallel} + \hat{z}\times\vec{g}_{np}$$
 (13)

Equation (13) is our final equation for the two dimensional current or  $\omega_z$  in the  $\hat{z}$  plane.

Roughly speaking, the terms in Equation (13), repeated below, have the following interpretation.

$$\vec{j}(\omega_z) = \vec{v}_{\parallel}\omega_z - v_z\vec{\omega}_{\parallel} + \hat{z} \times \vec{g}_{np}$$
 (13) repeated

The  $\vec{v}_{||}\omega_z$  term clearly represents the convection of  $\omega_z$  due to the fluid motion  $\vec{v}_{||}$  in the plane. The  $\hat{z} \times \vec{g}_{np}$  term which we call the *Magnus term*, gives us the sideways motion of the vortex when a non potential force is acting on the fluid. For example, if we have an  $\hat{x}$  directed force  $\vec{g}$  acting on the core of a  $\hat{z}$  directed vortex, we end up with a  $\hat{y}$  directed flow of vorticity as indicated in Figure (2), a diagram we have seen before.

The  $-v_z \vec{\omega}_{||}$  term is more of a problem to interpret. We note, however, that for a two dimensional flow with straight vortices, we can orient the  $\hat{z}$  plane to cut the vortex perpendicular to the core so that  $\vec{\omega}_{||}$  is zero and the term vanishes. We will see that for three dimensional fluid flow with a curved vortex, this term can be made to go away by choosing a properly oriented  $\hat{z}$  plane. From this point of view, the  $-v_z \vec{\omega}_{||}$  term tells us which  $\hat{z}$  plane to use.



**Figure 2** *Motion of a vortex line subject to an x directed force.* 

# A SINGLE VORTEX LINE

To help interpret the equations for vortex motion, we will apply Equation (13) to the motion of a single vortex line. We cut the line with a  $\hat{z}$  plane as shown in Figure (3a) and look at the behavior of  $\omega_z$  in that plane, as seen in Figure (3b). The main result is that we end up with a formula for the motion of the center of mass of  $\omega_z$ . This result is a consequence of their being a conserved two dimensional current of  $\omega_z$ .



**Figure 3a** *Cut the vortex line with a*  $\hat{z}$  *plane.* 



**Figure 3b** We will study the motion of  $\omega_{\tau}$  in the  $\hat{z}$  plane.

#### **Center of Mass Motion**

Our first step is to show that if we have an isolated vortex where both  $\omega_z$  and non potential forces  $\vec{g}_{np}$  are confined to a core region, then the vortex velocity  $\vec{V}_{vortex}$ , defined by

$$\vec{V}_{vortex} \equiv \frac{1}{\kappa} \int_{\substack{core \\ area}} \vec{j}(\omega_z) dA_z = \vec{V}_{COM}$$
 (14)

is the velocity of the center of mass of  $\omega_z$  in the  $\hat{z}$  plane.

To show this, we begin with Figure (4) where we show the localized core area of a vortex as it passes through the  $\hat{z}$  plane. We are assuming that the dotted rectangle from  $x_a$  to  $x_b$ , and  $y_a$  to  $y_b$  lies outside the core area where both  $\omega_z$  and  $\vec{j}(\omega_z)$  are zero.

We define the area  $\Delta A(y_i)$ , seen in Figure (4), as a band of thickness  $\Delta y$  that goes from  $x_a$  to  $x_b$ , and from  $y_i$  to  $y_i + \Delta y$ . The total vorticity  $\Delta \kappa_i$  in this band is

$$\Delta \kappa_{i} = \Delta y \int_{x_{a}}^{x_{b}} \omega_{z}(x, y_{i}) dx$$
(15)

The formula for the center of mass coordinate  $\vec{R}_{COM}$ of a collection of masses  $m_i$  is (see page 11-3 of the Physics text)

$$M\vec{R}_{COM} = \sum_{i} \vec{r}_{i}m_{i}$$
(16)

where M is the total mass.



Calculating the center of mass of  $\omega_7$ .

Replacing M by the vortex total circulation  $\kappa$ , and  $m_i$  by  $\Delta \kappa_i$ , the equation for the y component of the center of mass of the vorticity,  $Y_{COM}$ , becomes

$$\kappa Y_{\rm COM} = \sum_{i} y_i \Delta \kappa_i \tag{17}$$

Differentiating Equation (17) with respect to time, noting that the total circulation  $\kappa$  does not change with time, gives

$$\kappa \frac{\partial Y_{COM}}{\partial t} = \kappa V_{yCOM} = \sum_{i} y_{i} \frac{\partial \Delta \kappa_{i}}{\partial t}$$
(18)

Our problem now is to calculate the rate of change of the circulation  $\Delta \kappa_i$  in our  $\Delta y$  band. We do this by calculating the net rate of flow of vorticity into the band due to the vortex current  $\vec{j}(\omega_z)$ , indicated in Figure (5).

Along the line  $y = y_i$ , the net current into the band is

$$J_{y}(y_{i}) = \int_{x_{a}}^{x_{b}} j_{y}(x, y_{i}) dx \quad \stackrel{current in}{from \, below}$$
(19)

where  $J_y = J_y(\omega_z)$ .

Up at  $y_i + \Delta y$ , the component  $j_y(\omega_z)$  flows up out of the band, so that the net *inward* current up there has a minus sign

$$J_{y}(y_{i}+\Delta y)_{inward} = -J_{y}(y_{i}+\Delta y)$$
$$= -\int_{x_{a}}^{x_{b}} j_{y}(x, y_{i}+\Delta y) dx \qquad (20)$$



**Figure 5** *Flow of vorticity into band.* 

The total rate  $\partial \Delta \kappa_i / \partial t$  at which vorticity is flowing into the band is thus

$$\frac{\partial \Delta \kappa_i}{\partial t} = -J_y(y_i + \Delta y) + J_y(y_i)$$
(21)

Using Equation (21) in Equation (18) for  $V_{yCOM}$  gives

$$\kappa V_{yCOM} = -\sum_{i} y_{i} \Delta y \left[ \frac{J_{y}(y_{i} + \Delta y) - J_{y}(y_{i})}{\Delta y} \right] (22)$$

where we multiplied the right hand side by  $\Delta y / \Delta y$ .

In the limit  $\Delta y \rightarrow 0$ , the square brackets become the derivative  $\partial J_v(y)/\partial y$ , evaluated at  $y = y_i$ 

$$\kappa V_{yCOM} = -\sum_{i} y_{i} \left| \frac{\partial J_{y}(y)}{\partial y} \right|_{y=y_{i}} \left| \Delta y \right|$$
(23)

This sum  $\sum_{i} \Delta y$  then becomes an integral from  $y_a$  to  $y_b$ , giving i

$$\kappa V_{yCOM} = -\int_{y_a}^{y_b} y \frac{\partial J_y(y)}{\partial y} dy$$
(24)

The next step, which is called *integration by parts*, is a simple way to handle the factor y that appears in Equation (24). We note that, by the rules of differentiation

$$\frac{\partial}{\partial y} \left[ y J(y) \right] = \frac{\partial y}{\partial y} J(y) + y \frac{\partial J(y)}{\partial y}$$
(25)

With  $\partial y/\partial y = 1$  we get

$$y\frac{\partial J(y)}{\partial y} = \frac{\partial}{\partial y} [yJ(y)] - J(y)$$
(26)

Substituting (26) into (24) gives

$$\kappa V_{yCOM} = -\int_{y_a}^{y_b} \frac{\partial}{\partial y} \left[ y J_y(y) \right] dy + \int_{y_a}^{y_b} J_y(y) dy$$
(27)

We can explicitly carry out the first integral because the integral of a derivative is simply the function itself

$$-\int_{y_{a}}^{y_{b}} \frac{\partial}{\partial y} \left[ y J_{y}(y) \right] dy = y J_{y}(y) \Big|_{y_{a}}^{y_{b}}$$
$$= y_{b} J_{y}(y_{b}) - y_{a} J_{y}(y_{a}) \quad (28)$$
$$= 0$$

We get zero because both  $y_a$  and  $y_b$  lie outside the core region, where  $J_v$  is zero.

Thus we are left with

$$\kappa V_{yCOM} = \int_{y_a}^{y_b} J_y(y) dy$$
$$= \int_{y_a}^{y_b} \int_{x_a}^{x_b} j_y(x,y) dx dy$$
(29)

where we used Equation (19) to express  $J_y(y)$  in terms of the vortex current density  $j_y(x,y)$ .

Because we are assuming that  $j_y(x,y)$  is non zero only over the core area, Equation (29) can be written in the more compact form

$$\kappa V_{yCOM} = \int_{\substack{\text{core}\\ \text{area}}} j_y(\omega_z) dA_z$$
(30)

where  $dA_z = dx dy$ .

Similar arguments give

$$\kappa V_{\text{xCOM}} = \int_{\substack{\text{core}\\\text{area}}} j_{\text{x}}(\omega_{\text{z}}) dA_{\text{z}}$$
(31)

Combining Equation (30) and (31), and dividing through by  $\kappa$  gives

$$\vec{V}_{COM} = \frac{1}{\kappa} \int_{\substack{\text{core}\\\text{area}}} \vec{j}(\omega_z) dA_z \equiv \vec{V}_{\text{vortex}}$$
(14) repeate

(14) repeated

which is the result we wanted to show.

#### MAGNUS FORMULA FOR CURVED VORTICES

We are now ready to use Equation (13) to derive the Magnus effect formula for curved fluid core vortices. As a reminder, Equation (13) was

$$\vec{j}(\omega_z) = \vec{v}_{\parallel}\omega_z - v_z\vec{\omega}_{\parallel} + \hat{z} \times \vec{g}_{np}$$
 (13) repeated

Slicing a curved vortex with a  $\hat{z}$  plane as shown in Figure (3), integrating Equation (13) over the area of the core, and dividing through by  $\kappa$  gives

$$\frac{1}{\kappa} \int_{\substack{\text{core}\\\text{area}}} \vec{j}(\omega_z) dA_z = \frac{1}{\kappa} \int \omega_z \vec{v}_{\parallel} dA_z \qquad (32a)$$

+ 
$$\frac{1}{\kappa} \int -v_z \vec{\omega}_{\parallel} dA_z$$
 (32b)

+ 
$$\frac{1}{\kappa} \int \hat{z} \times \vec{g}_{np} dA_z$$
 (32c)

We already know that the left side of Equation (32) is the vortex velocity  $\vec{V}_{vortex}$ . The first term on the right, which we will call  $\vec{V}_{fluid}$ 

$$\vec{V}_{\text{fluid}} = \frac{1}{\kappa} \int \omega_z \vec{v}_{\parallel} dA_z$$
(33)

is the weighted average of the velocity field  $\vec{v}_{\parallel}$  in the core region.

As we mentioned earlier, the third term, the integral of  $v_z \vec{\omega}_{||}$  tells what  $\hat{z}$  plane to use for the calculation. There will be some plane, more or less perpendicular to the core, which gives a zero value for the integral of  $v_z \vec{\omega}_{||}$  over the core. We will assume that we are using that  $\hat{z}$  plane. For this example, let us assume that  $\vec{g}_{np}$  is an external force  $\vec{g}_e$  acting on the fluid in the core, as sketched in Figure (2) repeated below. Multiplying this force per unit mass by  $\rho$  gives  $\vec{f}_e = \rho \vec{g}_e$  as the force per unit volume acting on the core. When  $\vec{f}_e$  is integrated over the core, we get  $\vec{F}_e$ , the external force per unit length acting on the vortex.

With this notation the last term in Equation (32) becomes

$$\frac{1}{\kappa} \int_{\text{core}\atop\text{area}} \hat{z} \times \vec{g}_{np} dA_z = \frac{1}{\rho\kappa} \hat{z} \times \int_{\text{core}\atop\text{area}} \rho \vec{g}_{np} dA_z$$
$$= \frac{1}{\rho\kappa} \hat{z} \times \int_{\text{core}\atop\text{area}} \vec{f}_e dA_z$$
$$= \frac{\hat{z} \times \vec{F}_e}{\rho\kappa}$$
(34)

Assuming we have chosen the correct  $\hat{z}$  plane to eliminate the integral of  $v_z \vec{\omega}_{\parallel}$ , we get using Equations (14), (33) and (34) in Equation (32)

$$\vec{V}_{vortex} = \vec{V}_{fluid} + \frac{1}{\rho\kappa}\hat{z} \times \vec{F}_{e}$$
 (35)

The Helmholtz equation is now obtained by setting  $\vec{F}_e = 0$  giving

$$\vec{\mathbf{V}}_{\text{vortex}} = \vec{\mathbf{V}}_{\text{fluid}} \qquad \begin{array}{c} Helmholtz\\ equation for\\ \vec{F}_{o} = 0 \end{array} \tag{36}$$

In detail, Equation (36) says that when we choose the  $\hat{z}$  plane correctly, the center of mass motion of the vortex core is equal to the weighted average of the fluid velocity in the core region.



**Figure 2 (repeated)** *Motion of a vortex line subject to an x directed force.* 

When  $\vec{F}_e$  is not zero and we have a relative motion of the vortex line and the fluid, we can define the relative motion vector  $\vec{V}_{rel}$  as

$$\vec{V}_{rel} \equiv \vec{V}_{vortex} - \vec{V}_{fluid}$$
 (37)

and Equation (35) can be written

$$\vec{V}_{vortex} = \vec{V}_{fluid} + \frac{1}{\rho\kappa}\hat{z} \times \vec{F}_{e}$$
 (35) repeated

$$\hat{z} \times \vec{F}_e = \rho \kappa \vec{V}_{rel}$$
 (38)

We can get further insight from Equation (38) by writing  $\vec{F}_e$  as

$$\vec{F}_{e} = (\vec{F}_{ez} + \vec{F}_{e\perp})$$
 (39)

where  $\vec{F}_{ez}$  is the component of  $\vec{F}_e$  parallel to the  $\hat{z}$  axis, and  $\vec{F}_{e\perp}$  perpendicular to the z axis. Because  $\hat{z}$  cross a vector parallel to  $\hat{z}$  is zero,  $\hat{z} \times \vec{F}_{ez} = 0$  and we get

$$\hat{z} \times \vec{F}_{e} = \hat{z} \times \vec{F}_{e\perp} \tag{40}$$

Thus our final result for the Magnus equation is

$$\hat{z} \times \vec{F}_{e\perp} = \rho \kappa \vec{V}_{rel}$$
 Magnus  
equation (41)

and we see that only the component of the external force perpendicular to the z axis, has an effect on the vortex motion. This reminds us why it is important, for a curved vortex, to find the correct  $\hat{z}$  plane using the condition that the integral of  $v_z \vec{\omega}_{\parallel}$  be zero.

If we apply Equation (41) to a two dimensional flow in the xy plane, then the vorticity is automatically  $\hat{z}$ directed and we can turn  $\kappa$  into a  $\hat{z}$  directed vector  $\hat{\kappa}$ . If the flow is to remain two dimensional, then the external force  $\vec{F}_e$  must be in the xy plane, because a z component of  $\vec{F}_e$  would create a  $\hat{z}$  directed flow. Thus  $\vec{F}_e$  must be  $\vec{F}_{e\perp}$ . With these restrictions, Equation (33) is equivalent to

$$\vec{F}_{e} = \rho \vec{V}_{rel} \times \vec{\kappa}$$
(13-95)

which is our Equation (13-95) discussed in the regular part of the chapter. (Check for yourself that both Equations (41) and (13-95) predict that an  $\hat{x}$  directed force  $\vec{F}_e$  acting on a  $\hat{z}$  directed vortex causes a  $\hat{y}$  directed relative motion of the vortex.)

What we have learned from deriving the exact Magnus equation for curved vortices, that we cannot predict from a two dimensional derivation, is what component of  $\vec{F}_e$  is important and exactly how  $\vec{V}_{rel}$  is defined.

# **CREATION OF VORTICITY**

So far our emphasis has been on how non potential forces cause a relative motion of vortex lines and the fluid particles. But the vorticity we find in a fluid has to have been created somehow. Non potential forces do that, and we want to end this appendix with a brief discussion of how. The discussion is brief, because it is very incomplete. The creation of vorticity, which leads to turbulence, is not only a subject for an entire fluid dynamics textbook, it is also an active subject of current research. Here we will just indicate how the topic begins.

Non potential forces, at least in a constant density fluid like water, can create vorticity in two ways. One way is to pull it out of the walls of the container. Near the wall, where the velocity field rapidly goes to zero, we get a boundary layer where the non potential viscous forces are important. These viscous forces, if they are acting at the wall, will move vorticity out of the wall into the fluid. For example, this is how the vorticity in the smoke ring demonstration was created. Viscous forces acting on the high speed fluid at the perimeter of the hole in the box pulled a ring of vorticity in from the perimeter.

It turns out to be a tricky question of how viscous forces behave in a boundary layer. For laminar pipe flow, there are viscous forces acting at the wall continually pulling vorticity into the stream. In contrast, for a boundary layer solution called the *Blasius profile*, the viscous forces act in the boundary layer but not at the wall. In that theory, the vorticity is all created upstream and all the viscous forces do is move the vorticity farther into the fluid, thickening the boundary layer. The velocity profiles near the wall look nearly the same for both laminar pipe flow and the Blasius profile, but the viscous effects are quite different. This indicates the kind of problem one has to deal with when working with boundary layers and the effects of viscosity.

Non potential forces can also create vorticity in the fluid away from the walls by creating vortex rings. In a sense, this is the way vorticity is created in the Rayfield-Reif experiment. To give you a rough classical picture of how a charged particle moving through a fluid could create a vortex ring, imagine that the charged particle, moving in what we will call the  $\hat{z}$  direction exerts a local, more or less spherical shaped external force  $\vec{g}$  on the fluid as shown in Figure (5). This looks much like the figure we have drawn so many times, except that there is no vortex line for  $\vec{g}$  to push on. Thus  $\vec{g}$  cannot be causing a relative motion of the line and the fluid. What it is doing instead is creating a vortex ring around the region.

We can see the ring creation by applying the extended Helmholtz equation (12-78) to the circuits  $C'_1$ ,  $C'_2$  and  $C'_3$  shown in Figure (6). These circuits are moving with the fluid particles, and Equation (78) tells us that the rate of change of flux of  $\omega$  through any of them is equal to  $\oint \vec{g} \cdot d\vec{\ell}$  around the circuit. With this in mind, we see that the flux of  $\omega$  through  $C'_1$  is increasing because  $\oint \vec{g} \cdot d\vec{\ell}$  is positive there, and it is decreasing through  $C'_2$  where  $\oint \vec{g} \cdot d\vec{\ell}$  is negative. Since  $\oint \vec{g} \cdot d\vec{\ell}$  is zero for  $C'_3$ , there is no change in the flux of  $\omega$  there.

What does it mean that  $\vec{g}$  is decreasing the flux through the lower circuit  $C'_2$  when there is no flux there to decrease? It means that  $\vec{g}$  is creating negative flux of  $\vec{\omega}$  through  $C'_2$  while at the same time it is creating positive flux through  $C'_1$ . What it is doing is creating a *band of flux* of  $\vec{\omega}$  around the spherical region, a band of flux that is becoming the core of a vortex ring.

Once vorticity has been introduced into the fluid, an effective method of introducing more vorticity is the stretching of existing vortex lines. How vortex line stretching affects fluid flows is a topic that has been studied for a long time by fluid engineers.



**Figure 6** *External force creating a vortex ring.*
# **ENERGY DISSIPATION IN FLUID FLOW**

While a derivation of the Magnus formula for curved vortices demonstrates how mathematically effective the concept of a vortex current  $\vec{j}(\omega_z)$  is, (the result has not been obtained any other way), the most important use so far of the concept is in studying the relationship between energy dissipation in a stream and the flow of vorticity across the stream. This relationship, discovered by Phillip Anderson in 1966, applies to such diverse situations as turbulent flow in a channel, and the motion of quantized vortices in both superfluids and superconductors. In the case of superconductors, the phenomenon is now involved in the legal definition of the electric volt.

We leave this topic for a later text, because one of the most interesting parts is to show how similar the vortex dynamics equations are for charged and neutral fluids. One can make the equations look identical by incorporating the magnetic field  $\vec{B}$  in the definition of  $\vec{\omega}$ , and including the electric field  $\vec{E}$  in  $\vec{g}_{np}$ . If you want to see this topic now, look at the article "Vortex Currents in Turbulent Superfluid and Classical Fluid Channel Flow . . .", Huggins, E.R., *Journal of Low Temperature Physics*, Vol. 96, 1994.

The 1852 article by Magnus is "On the deviation of projectiles; and on a remarkable phenomenon of rotating bodies." G. Magnus, Memoirs of the Royal Academy, Berlin (1852). English translation in Scientific Memoirs, London (1853), p.210. Edited by John Tyndall and William Francis.

# Formulary For Vector Operations

When you are working problems involving quantities like  $\nabla^2$  in cylindrical or spherical coordinates, you do not want to derive the formulas yourself because the chances of your getting the right answer are too small. You are not likely to memorize them correctly either, unless you use a particular formula often. Instead, the best procedure is to look up the result in a table of formulas, sometimes called a **formulary**.

In this formulary we summarize all the formulas for gradient, divergence and curl, in Cartesian, cylindrical and spherical coordinates. We also include integral formulas, formulas for working with cross products, and with tensors. The formulary was adapted from one developed by David Book of the Naval Research Laboratory.

We have also added a short table of integrals, and summarize some of the series expansions we discussed in the text.

# Contents

Cylindrical Coordinates	Formulary - 2
Divergence	Formulary - 2
Gradient	Formulary - 2
Curl	Formulary - 2
Laplacian	Formulary - 2
Laplacian of a vector	Formulary - 2
Components of $(\vec{A}\cdot\vec{\nabla})$ $\vec{B}$	Formulary - 2
Spherical Polar Coordinates	Formulary - 3
Divergence	Formulary - 3
Gradient	Formulary - 3
Curl	Formulary - 3
Laplacian	Formulary - 3
Laplacian of a vector	Formulary - 3
Components of (A · V) B	Formulary - 3
Vector Identities	Formulary - 4
Integral Formulas	Formulary - 5
Working with Cross Products	Formulary - 6
The Cross Product	Formulary - 6
Product of e's	Formulary - 6
Example of use	Formulary - 6
Tensor Formulas	Formulary - 7
Definition	Formulary - 7
Formulas	Formulary - 7
Div. (Cylindrical Coord.)	Formulary - 7
Div. (Spherical Coord.)	Formulary - 7
Short Table of Integrals	Formulary - 8
Series Expansions	Formulary - 9
The Binomial Expansion	Formulary - 9
Taylor Series Expansion	Formulary - 9
Sine and Cosine	Formulary - 9
Exponential	Formulary - 9

# **CYLINDRICAL COORDINATES**

# Divergence

$$\vec{\nabla}\cdot\vec{A} \;=\; \frac{1}{r}\;\frac{\partial}{\partial r}\left(rA_{r}\right) + \frac{1}{r}\;\frac{\partial A_{\theta}}{\partial \theta} + \frac{\partial A_{z}}{\partial z}$$

# Gradient

$$(\vec{\nabla}f)_{r} = \frac{\partial f}{\partial r}$$
$$(\vec{\nabla}f)_{\theta} = \frac{1}{r} \frac{\partial f}{\partial \theta}$$
$$(\vec{\nabla}f)_{z} = \frac{\partial f}{\partial z}$$

### Curl

$$(\vec{\nabla} \times \vec{A})_{r} = \frac{1}{r} \frac{\partial A_{z}}{\partial \theta} - \frac{\partial A_{\theta}}{\partial z}$$
$$(\vec{\nabla} \times \vec{A})_{\theta} = \frac{\partial A_{r}}{\partial z} - \frac{\partial A_{z}}{\partial r}$$
$$(\vec{\nabla} \times \vec{A})_{z} = \frac{1}{r} \frac{\partial}{\partial r} (rA_{\theta}) - \frac{1}{r} \frac{\partial A_{r}}{\partial \theta}$$

# Laplacian

$$\nabla^{2} \mathbf{f} = \frac{1}{\mathbf{r}} \frac{\partial}{\partial \mathbf{r}} (\mathbf{r} \frac{\partial \mathbf{f}}{\partial \mathbf{r}}) + \frac{1}{\mathbf{r}^{2}} \frac{\partial^{2} \mathbf{f}}{\partial \theta^{2}} + \frac{\partial^{2} \mathbf{f}}{\partial z^{2}}$$

# Laplacian of a vector

$$(\nabla^{2}\vec{A})_{r} = \nabla^{2}A_{r} - \frac{2}{r^{2}}\frac{\partial A_{\theta}}{\partial \theta} - \frac{A_{r}}{r^{2}}$$
$$(\nabla^{2}\vec{A})_{\theta} = \nabla^{2}A_{\theta} + \frac{2}{r^{2}}\frac{\partial A_{r}}{\partial \theta} - \frac{A_{\theta}}{r^{2}}$$
$$(\nabla^{2}\vec{A})_{z} = \nabla^{2}A_{z}$$

# Components of $(\vec{A} \cdot \vec{\nabla}) \vec{B}$





**Cartesian Coordinates** 



**Cylindrical Coordinates** 

# **SPHERICAL POLAR COORDINATES**

# Divergence

$$\vec{\nabla} \cdot \vec{A} = \frac{1}{r^2} \frac{\partial}{\partial r} (r^2 A_r) + \frac{1}{r \sin \theta} \frac{\partial}{\partial \theta} (A_\theta \sin \theta) + \frac{1}{r \sin \theta} \frac{\partial A_\phi}{\partial \phi}$$

# Gradient

$$(\vec{\nabla}f)_{r} = \frac{\partial f}{\partial r}$$
$$(\vec{\nabla}f)_{\theta} = \frac{1}{r}\frac{\partial f}{\partial \theta}$$
$$(\vec{\nabla}f)_{\phi} = \frac{1}{r\sin\theta}\frac{\partial f}{\partial \phi}$$

# 

**Spherical Polar Coordinates** 

Ζ

# Curl

$$(\vec{\nabla} \times \vec{A})_{r} = \frac{1}{r \sin\theta} \frac{\partial}{\partial \theta} (A_{\phi} \sin\theta) - \frac{1}{r \sin\theta} \frac{\partial A_{\theta}}{\partial \phi}$$
$$(\vec{\nabla} \times \vec{A})_{\theta} = \frac{1}{r \sin\theta} \frac{\partial A_{r}}{\partial \phi} - \frac{1}{r} \frac{\partial}{\partial r} (rA_{\phi})$$
$$(\vec{\nabla} \times \vec{A})_{\phi} = \frac{1}{r} \frac{\partial}{\partial r} (rA_{\theta}) - \frac{1}{r} \frac{\partial A_{r}}{\partial \theta}$$

# Laplacian

$$\nabla^{2} f = \frac{1}{r} \frac{\partial^{2}}{\partial r^{2}} (rf) + \frac{1}{r^{2} \sin \theta} \frac{\partial}{\partial \theta} \left( \sin \theta \frac{\partial f}{\partial \theta} \right) + \frac{1}{r^{2} \sin^{2} \theta} \frac{\partial^{2} f}{\partial \phi^{2}}$$

# Laplacian of a vector

$$\begin{aligned} (\nabla^2 \vec{A})_r &= \nabla^2 A_r - \frac{2}{r^2} \frac{\partial A_\theta}{\partial \theta} - \frac{2A_\theta \cot\theta}{r^2} - \frac{2}{r^2 \sin\theta} \frac{\partial A_\phi}{\partial \phi} \\ (\nabla^2 \vec{A})_\theta &= \nabla^2 A_\theta + \frac{2}{r^2} \frac{\partial A_r}{\partial \theta} - \frac{A_\theta}{r^2 \sin^2 \theta} - \frac{2\cos\theta}{r^2 \sin^2 \theta} \frac{\partial A_\phi}{\partial \phi} \\ (\nabla^2 \vec{A})_\phi &= \nabla^2 A_\phi - \frac{A_\phi}{r^2 \sin^2 \theta} + \frac{2}{r^2 \sin\theta} \frac{\partial A_r}{\partial \phi} + \frac{2\cos\theta}{r^2 \sin^2 \theta} \frac{\partial A_\phi}{\partial \phi} \end{aligned}$$

# Components of $(\vec{A} \cdot \vec{\nabla}) \vec{B}$

$$\begin{split} & [(\vec{A} \cdot \vec{\nabla})\vec{B}]_{r} = A_{r} \frac{\partial B_{r}}{\partial r} + \frac{A_{\theta}}{r} \frac{\partial B_{r}}{\partial \theta} + \frac{A_{\phi}}{r \sin \theta} \frac{\partial B_{r}}{\partial \phi} - \frac{A_{\theta}B_{\theta} + A_{\phi}B_{\phi}}{r} \\ & [(\vec{A} \cdot \vec{\nabla})\vec{B}]_{\theta} = A_{r} \frac{\partial B_{\theta}}{\partial r} + \frac{A_{\theta}}{r} \frac{\partial B_{\theta}}{\partial \theta} + \frac{A_{\phi}}{r \sin \theta} \frac{\partial B_{\theta}}{\partial \phi} + \frac{A_{\theta}B_{r}}{r} - \frac{A_{\phi}B_{\phi}cot\theta}{r} \\ & [(\vec{A} \cdot \vec{\nabla})\vec{B}]_{\phi} = A_{r} \frac{\partial B_{\phi}}{\partial r} + \frac{A_{\theta}}{r} \frac{\partial B_{\phi}}{\partial \theta} + \frac{A_{\phi}}{r \sin \theta} \frac{\partial B_{\phi}}{\partial \phi} + \frac{A_{\phi}B_{r}}{r} - \frac{A_{\phi}B_{\phi}cot\theta}{r} \end{split}$$

Formulary-4

# **VECTOR IDENTITIES**

Notation: f, g, etc., are scalars;  $\vec{A}\,$  and  $\vec{B}\,,$  etc. are vectors

(1) 
$$\vec{A} \cdot \vec{B} \times \vec{C} = \vec{A} \times \vec{B} \cdot \vec{C} = \vec{B} \cdot \vec{C} \times \vec{A} = \vec{B} \times \vec{C} \cdot \vec{A} = \vec{C} \cdot \vec{A} \times \vec{B} = \vec{C} \times \vec{A} \cdot \vec{B}$$
  
(2)  $\vec{A} \times (\vec{B} \times \vec{C}) = (\vec{A} \cdot \vec{C}) \vec{B} - (\vec{A} \cdot \vec{B}) \vec{C}$   
(3)  $\vec{A} \times (\vec{B} \times \vec{C}) + \vec{B} \times (\vec{C} \times \vec{A}) + \vec{C} \times (\vec{A} \times \vec{B}) = 0$   
(4)  $(\vec{A} \times \vec{B}) \cdot (\vec{C} \times \vec{D}) = (\vec{A} \cdot \vec{C}) (\vec{B} \cdot \vec{D}) - (\vec{A} \cdot \vec{D}) (\vec{B} \cdot \vec{C})$   
(5)  $(\vec{A} \times \vec{B}) \times (\vec{C} \times \vec{D}) = (\vec{A} \times \vec{B} \cdot \vec{D}) \vec{C} - (\vec{A} \times \vec{B} \cdot \vec{C}) \vec{D}$   
(6)  $\vec{\nabla} (fg) = \vec{\nabla} (gf) = f \vec{\nabla} (g) + g \vec{\nabla} (f)$   
(7)  $\vec{\nabla} \cdot (\vec{A}) = f \vec{\nabla} \cdot \vec{A} + \vec{A} \cdot \vec{\nabla} f$   
(8)  $\vec{\nabla} \times (\vec{A}) = f \vec{\nabla} \times \vec{A} + \vec{\nabla} f \times \vec{A}$   
(9)  $\vec{\nabla} \cdot (\vec{A} \times \vec{B}) = \vec{B} \cdot \vec{\nabla} \times \vec{A} - \vec{A} \cdot \vec{\nabla} \times \vec{B}$   
(10)  $\vec{\nabla} \times (\vec{A} \times \vec{B}) = \vec{A} \cdot (\vec{\nabla} \times \vec{B}) - \vec{B} (\vec{\nabla} \cdot \vec{A}) + (\vec{B} \cdot \vec{\nabla}) \vec{A} - (\vec{A} \cdot \vec{\nabla}) \vec{B}$   
(11)  $\vec{\nabla} (\vec{A} \cdot \vec{B}) = \vec{A} \times (\vec{\nabla} \times \vec{B}) + \vec{B} \times (\vec{\nabla} \times \vec{A}) + (\vec{A} \cdot \vec{\nabla}) \vec{B} + (\vec{B} \cdot \vec{\nabla}) \vec{A}$   
(12)  $\nabla^2 f = \vec{\nabla} \cdot \vec{\nabla} f$   
(13)  $\nabla^2 \vec{A} = \vec{\nabla} (\vec{\nabla} \cdot \vec{A}) - \vec{\nabla} \times \vec{\nabla} \times \vec{A}$   
 $\vec{\nabla} \times (\vec{\nabla} \times \vec{A}) = \vec{\nabla} (\vec{\nabla} \cdot \vec{A}) - \nabla^2 \vec{A}$   
(14)  $\vec{\nabla} \times \vec{\nabla} f = 0$   
(15)  $\vec{\nabla} \cdot \vec{\nabla} \times \vec{A} = 0$   
Let  $\vec{r} = \vec{i} x + \vec{j} y + \vec{k} z$  be the radius vector of magnitude r, from the origin to the point x, y, z. Then  
(16)  $\vec{\nabla} \cdot \vec{r} = 3$   
(17)  $\vec{\nabla} \times \vec{r} = 0$   
(18)  $\vec{\nabla} r = \vec{r}/r$   
(19)  $\vec{\nabla} (1/r) = -\vec{r}/r^3$   
(20)  $\vec{\nabla} \cdot (\vec{r}/r^3) = 4\pi \delta(\vec{r})$ 

# **INTEGRAL FORMULAS**

If V is the volume enclosed by a surface S and  $\vec{dS} = \hat{n}dS$  where  $\hat{n}$  is the unit normal outward from V

$$(22) \int_{V} \vec{\nabla} f \, d^{3} V = \int_{S} f \, \vec{dS}$$

$$(23) \int_{V} \vec{\nabla} \cdot \vec{A} \, d^{3} V = \int_{S} \vec{A} \cdot \vec{dS}$$

$$(24) \int_{V} \vec{\nabla} \times \vec{A} \, d^{3} V = \int_{S} \vec{dS} \times \vec{A}$$

$$(25) \int_{V} (f \nabla^{2} g - g \nabla^{2} f) \, d^{3} V = \int_{S} (f \vec{\nabla} g - g \vec{\nabla} f) \cdot \vec{dS}$$

$$(26) \int_{V} \left[ \vec{A} \cdot \vec{\nabla} \times (\vec{\nabla} \times \vec{B}) - \vec{B} \cdot \vec{\nabla} \times (\vec{\nabla} \times \vec{A}) \right] d^{3} V = \int_{S} \left[ \vec{B} \times (\vec{\nabla} \times \vec{A}) - \vec{A} \times (\vec{\nabla} \times \vec{B}) \right] \cdot \vec{dS}$$

If S is an open surface bounded by the contour C of which the line element is  $\vec{d\ell}$ 

$$(27) \int_{S} \vec{dS} \times \vec{\nabla} f = \oint_{C} \vec{fd\ell}$$

$$(28) \int_{S} (\vec{\nabla} \times \vec{A}) \cdot \vec{dS} = \oint_{C} \vec{A} \cdot \vec{d\ell} \qquad \begin{array}{l} Stokes'\\ law \end{array}$$

$$(29) \int_{S} (\vec{dS} \times \vec{\nabla}) \times \vec{A} = \oint_{C} \vec{d\ell} \times \vec{A}$$

$$(30) \int_{S} (\vec{\nabla} f \times \vec{\nabla} g) \cdot \vec{dS} = \oint_{C} \vec{fdg} = -\oint_{C} \vec{g}df$$

# **WORKING WITH CROSS PRODUCTS**

Use of the permutation tensor  $\epsilon_{ijk}$  to work effectively with the cross products. (Reference: Appendix I in Chapter 13.)

# The cross product

$$(\vec{A} \times \vec{B})_i = \varepsilon_{ijk} A_j B_k$$

# Product of $\epsilon$ 's

 $\epsilon_{ijk}\epsilon_{klm}\ =\ \delta_{il}\delta_{jm}-\delta_{im}\delta_{jl}$ 

# Example of use

$$\begin{bmatrix} \vec{\nabla} \times (\vec{\nabla} \times \vec{A}) \end{bmatrix}_{i} = \epsilon_{ijk} \nabla_{j} (\vec{\nabla} \times \vec{A})_{k}$$
$$= \epsilon_{ijk} \epsilon_{klm} \nabla_{j} \nabla_{l} A_{m}$$
$$= (\delta_{il} \delta_{jm} - \delta_{im} \delta_{jl}) \nabla_{j} \nabla_{l} A_{m}$$
$$= \nabla_{j} \nabla_{i} A_{j} - \nabla_{j} \nabla_{j} A_{i}$$
$$= \nabla_{i} \nabla_{j} A_{j} - \nabla_{j} \nabla_{j} A_{i}$$
$$= \vec{\nabla} (\vec{\nabla} \cdot \vec{A}) - \nabla^{2} \vec{A}$$

# **TENSOR FORMULAS**

Notation: f, g, etc., are scalars;  $\vec{A}$  and  $\vec{B}$ , etc. are vectors;  $\vec{T}$  is a tensor

# Definition

If  $\vec{e}_1$ ,  $\vec{e}_2$ ,  $\vec{e}_3$  are orthonormal unit vectors, a second-order tensor  $\vec{T}$  can be written in the dyadic form

$$\vec{\bar{T}} = \sum_{i,j} T_{ij} \vec{e}_i \vec{e}_j$$

In Cartesian coordinates the divergence of a tensor is a vector with components

$$(\nabla\cdot\vec{\vec{T}})_{i} = \sum_{j} (\partial T_{ji}/\partial x_{j})$$

# Formulas

$$\vec{\nabla} \cdot (\vec{A}\vec{B}) = (\vec{\nabla} \cdot \vec{A})\vec{B} + (\vec{A} \cdot \vec{\nabla})\vec{B}$$
$$\vec{\nabla} \cdot (\vec{f}\vec{T}) = \vec{\nabla}f \cdot \vec{T} + \vec{f}\vec{\nabla} \cdot \vec{T}$$
$$\int_{V} \vec{\nabla} \cdot \vec{T} d^{3}V = \int_{S} \vec{dS} \cdot \vec{T}$$

# **Divergence of a tensor (cylindrical coordinates)**

$$\begin{split} (\vec{\nabla} \cdot \vec{\vec{T}})_{r} &= \frac{1}{r} \frac{\partial}{\partial r} (rT_{rr}) + \frac{1}{r} \frac{\partial}{\partial \theta} (T_{\theta r}) + \frac{\partial T_{zr}}{\partial z} - \frac{1}{r} T_{\theta \theta} \\ (\vec{\nabla} \cdot \vec{\vec{T}})_{\theta} &= \frac{1}{r} \frac{\partial}{\partial r} (rT_{r\theta}) + \frac{1}{r} \frac{\partial T_{\theta \theta}}{\partial \theta} + \frac{\partial T_{z\theta}}{\partial z} + \frac{1}{r} T_{\theta r} \\ (\vec{\nabla} \cdot \vec{\vec{T}})_{z} &= \frac{1}{r} \frac{\partial}{\partial r} (rT_{rz}) + \frac{1}{r} \frac{\partial T_{\theta z}}{\partial \theta} + \frac{\partial T_{zz}}{\partial z} \end{split}$$

# **Divergence of a tensor (spherical coordinates)**

$$\begin{aligned} (\vec{\nabla} \cdot \vec{\vec{T}})_{r} &= \frac{1}{r^{2}} \frac{\partial}{\partial r} (r^{2} T_{rr}) + \frac{1}{r \sin \theta} \frac{\partial}{\partial \theta} (T_{\theta r} \sin \theta) + \frac{1}{r \sin \theta} \frac{\partial T_{\phi r}}{\partial \phi} - \frac{T_{\theta \theta} + T_{\phi \phi}}{r} \\ (\vec{\nabla} \cdot \vec{\vec{T}})_{\theta} &= \frac{1}{r^{2}} \frac{\partial}{\partial r} (r^{2} T_{r\theta}) + \frac{1}{r \sin \theta} \frac{\partial}{\partial \theta} (T_{\theta \theta} \sin \theta) + \frac{1}{r \sin \theta} \frac{\partial T_{\phi \theta}}{\partial \phi} + \frac{T_{\theta r}}{r} - \frac{\cot \theta}{r} T_{\phi \phi} \\ (\vec{\nabla} \cdot \vec{\vec{T}})_{\phi} &= \frac{1}{r^{2}} \frac{\partial}{\partial r} (r^{2} T_{r\phi}) + \frac{1}{r \sin \theta} \frac{\partial}{\partial \theta} (T_{\theta \phi} \sin \theta) + \frac{1}{r \sin \theta} \frac{\partial T_{\phi \phi}}{\partial \phi} + \frac{T_{\phi r}}{r} + \frac{\cot \theta}{r} T_{\phi \theta} \end{aligned}$$

# SHORT TABLE OF INTEGRALS

In these integrals, (a) is a constant, and (u) and (v) are any functions of x.

$$1. \int dx = x \qquad 10. \int \sin^{2}x \, dx = \frac{1}{2}x - \frac{1}{4}\sin 2x$$

$$2. \int au \, dx = a \int u \, dx \qquad 11. \int e^{-ax} dx = -\frac{1}{a}e^{-ax}$$

$$3. \int (u + v) \, dx = \int u \, dx + \int v \, dx \qquad 12. \int xe^{-ax} dx = -\frac{1}{a^{2}}(ax + 1)e^{-ax}$$

$$4. \int x^{m} \, dx = \frac{x^{m+1}}{m+1}(m \neq -1) \qquad 13. \int x^{2}e^{-ax} dx = -\frac{1}{a^{3}}(a^{2}x^{2} + 2ax + 2)e^{-ax}$$

$$5. \int \frac{dx}{x} = \ln |x| \qquad 14. \int_{0}^{\infty} x^{n}e^{-ax} dx = \frac{n!}{a^{n+1}}$$

$$6. \int u \, \frac{dv}{dx} dx = uv - \int v \, \frac{du}{dx} dx \qquad 15. \int_{0}^{\infty} x^{2n}e^{-ax^{2}} dx = \frac{1 \cdot 3 \cdot 5 \cdots (2n-1)}{2^{n+1}a^{n}} \sqrt{\frac{\pi}{a}}$$

$$8. \int \sin x \, dx = -\cos x \qquad 16. \int \frac{dx}{(x^{2} + a^{2})^{3/2}} = \frac{x}{a^{2}\sqrt{x^{2} + a^{2}}}$$

### SERIES EXPANSIONS

### The binomial expansion

(Ch 2, page 6)

$$(1 + \alpha)^n = 1 + n\alpha + \frac{n(n-1)}{2!}\alpha^2 + \cdots$$
 (2-22)

which is valid for any value of  $\alpha$  less than one, but which gets better as  $\alpha$  becomes smaller.

### **Taylor series expansion**

(Ch 2, page 8)

$$f(x - x_0) = f(x_0) + f'(x_0)(x - x_0)^1 + \frac{1}{2!}f''(x_0)(x - x_0)^2 + \frac{1}{3!}f'''(x_0)(x - x_0)^3 + \cdots$$

This can be written in the compact form

$$f(x - x_0) = \sum_{n=0}^{\infty} \frac{f^n(x_0)}{n!} (x - x_0)^n \qquad \begin{array}{c} Taylor \\ series \\ expansion \end{array}$$

(2-44)

•

where we used the notation

$$f^{n}(x_{0}) \equiv \left. \frac{d^{n}f(x)}{dx^{n}} \right|_{x = x_{0}}$$
 (2-45)

### Sine and cosine

(Ch 5, page 4)

$$\cos \theta = 1 - \frac{\theta^2}{2!} + \frac{\theta^4}{4!} + \cdots$$
 (13)

$$\sin \theta = \theta - \frac{\theta^3}{3!} + \frac{\theta^5}{5!} + \cdots$$
 (14)

where  $\theta$  is in radians. These expansions are valid for any value of  $\theta$ , but most useful for small values where we do not have to keep many terms.

### Exponential

(Ch 1, page 28 and Ch 5, page 4)

$$e^{x} = 1 + x + \frac{x^{2}}{2!} + \frac{x^{3}}{3!} + \cdots$$
 (1-136)

While this expansion is true for any value of x, it is most useful for small values of x where we do not have to keep many terms to get an accurate answer.

Setting  $x = i\theta$  gives

$$e^{i\theta} = 1 + i\theta + \frac{i^2\theta^2}{2!} + \frac{i^3\theta^3}{3!} + \cdots$$
 (5-12)

(Since our previous discussion of exponents only dealt with real numbers, we can consider Equation (12) as the definition of what we mean when the exponent is a complex number).

# **Physical Constants in CGS Units**

speed of light	$c = 3 \times 10^{10} \text{ cm/sec} = 1000 \text{ ft/}\mu\text{sec} = 1 \text{ ft/}nanosecond$
acceleration due to gravity	
at the surface of the earth	$g = 980 \text{ cm/sec}^2 = 32 \text{ ft/sec}^2$
gravitational constant	$G = 6.67 \times 10^{-8} \text{ cm}^3 / (\text{gm sec}^2)$
charge on an electron	$e = 4.8 \times 10^{-10} esu$
Planck's constant	h = $6.62 \times 10^{-27}$ erg sec (gm cm <sup>2</sup> /sec)
Planck constant / $2\pi$	$\hbar = 1.06 \times 10^{-27} \text{erg sec} (\text{gm cm}^2 / \text{sec})$
Bohr radius	$a_0 = .529 \times 10^{-8} cm$
rest mass of electron	$m_e = 0.911 \times 10^{-27} gm$
rest mass of proton	$M_{\rm D} = 1.67 \times 10^{-24} {\rm gm}$
rest energy of electron	$m_{e}^{r}c^{2} = 0.51 \text{ MeV} (\approx 1/2 \text{ MeV})$
rest energy of proton	$M_{p}c^{2} = 0.938 \text{ BeV} (\approx 1 \text{ BeV})$
proton radius	$r_{p} = 1.0 \times 10^{-13} cm$
Boltzmann's constant	$\dot{k} = 1.38 \times 10^{-16} \text{ergs/kelvin}$
Avogadro's number	$N_0 = 6.02 \times 10^{-23}$ molecules/mole

. - 10

. . . . . .

absolute zero =  $0^{\circ}$ K =  $-273^{\circ}$ C density of mercury =  $13.6 \text{ gm} / \text{cm}^3$ mass of earth =  $5.98 \times 10^{27}$  gm mass of the moon =  $7.35 \times 10^{25}$  gm mass of the sun =  $1.97 \times 10^{33}$  gm earth radius =  $6.38 \times 10^{8}$  cm = 3960 mi moon radius =  $1.74 \times 10^{8}$  cm = 1080 mi mean distance to moon =  $3.84 \times 10^{10}$  cm mean distance to sun =  $1.50 \times 10^{13}$  cm mean earth velocity in orbit about sun = 29.77 km / sec

### **Conversion Factors**

1 meter = 100 cm (100 cm/meter) 1 in. = 2.54 cm (2.54 cm/in.)1 mi = 5280 ft (5280 ft/mi)  $1 \text{ km} (\text{kilometer}) = 10^5 \text{ cm} (10^5 \text{ cm} / \text{ km})$  $1 \text{ mi} = 1.61 \text{ km} = 1.61 \times 10^5 \text{ cm} (1.61 \times 10^5 \text{ cm/mi})$ 1 Å (angstrom) =  $10^{-8}$  cm ( $10^{-8}$  cm / Å)  $1 \text{ day} = 86,000 \text{ sec} (8.6 \times 10^4 \text{ sec} / \text{ day})$ 1 year =  $3.16 \times 10^7$  sec (3.16×10<sup>7</sup> sec/year)  $1 \mu \text{sec} (\text{microsecond}) = 10^{-6} \text{sec} (10^{-6} \text{sec} / \mu \text{sec})$ 1 nanosecond =  $10^{-9} \text{sec} (10^{-9} \text{sec} / \text{nanosecond})$ 1 mi/hr = 44.7 cm/sec60 mi/hr = 88 ft/sec  $1 \text{ kg} (\text{kilogram}) = 10^{3} \text{gm} (10^{3} \text{gm} / \text{kg})$ 1 coulomb =  $3 \times 10^9$  esu ( $3 \times 10^9$  esu/coulomb) 1 ampere =  $3 \times 10^9$  statamps ( $3 \times 10^9$  statamps/ampere) 1 statvolt = 300 volts (300 volts/statvolt)  $1 \text{ joule} = 10^7 \text{ ergs} (10^7 \text{ ergs} / \text{ joule})$  $1 \text{ W} (\text{watt}) = 10^7 \text{ ergs/sec} (10^7 \text{ erg}/\text{W})$  $1 \text{ eV} = 1.6 \times 10^{-12} \text{ ergs}$  ( $1.6 \times 10^{-12} \text{ ergs/eV}$ )  $1 \text{ MeV} = 10^{6} \text{eV} (10^{6} \text{eV}/\text{MeV})$  $1 \text{ BeV} = 10^9 \text{eV} (10^9 \text{eV}/\text{BeV})$  $1 \mu$  (micron) pressure = 1.33 dynes / cm<sup>2</sup> 1 cm Hg pressure =  $10^4 \mu$  $1 \text{ atm} = 76 \text{ cm Hg} = 1.01 \times 10^{6} \text{ dynes/cm}^{2}$ 



 $\overline{\phantom{a}}$ 

Moose Mountain Digital Press