

Sherlock Holmes in Babylon

and Other Tales of
Mathematical History

Editors

Marlow Anderson,
Victor Katz,
Robin Wilson



The Mathematical Association of America

Sherlock Holmes in Babylon
and Other Tales of Mathematical History

© 2004 by
The Mathematical Association of America (Incorporated)
Library of Congress Catalog Card Number 2003113541

Print ISBN: 978-0-88385-546-1

Electronic ISBN: 978-1-61444-503-6

Printed in the United States of America

Current Printing (last digit):

10 9 8 7 6 5 4 3 2

Sherlock Holmes in Babylon

and Other Tales of Mathematical History

Edited by

Marlow Anderson

Colorado College

Victor Katz

University of the District of Columbia

Robin Wilson

Open University



Published and Distributed by
The Mathematical Association of America

Committee on Publications

Gerald L. Alexanderson, *Chair*

Spectrum Editorial Board

Gerald L. Alexanderson, *Chair*

Robert Beezer	Russell L. Merris
William Dunham	Jean J. Pedersen
Michael Filaseta	J. D. Phillips
Erica Flapan	Marvin Schaefer
Eleanor Lang Kendrick	Harvey Schmidt
Jeffrey L. Nunemacher	Sanford Segal
Ellen Maycock	Franklin Sheehan
John E. Wetzel	

SPECTRUM SERIES

The Spectrum Series of the Mathematical Association of America was so named to reflect its purpose: to publish a broad range of books including biographies, accessible expositions of old or new mathematical ideas, reprints and revisions of excellent out-of-print books, popular works, and other monographs of high interest that will appeal to a broad range of readers, including students and teachers of mathematics, mathematical amateurs, and researchers.

777 Mathematical Conversation Starters, by John de Pillis

All the Math That's Fit to Print, by Keith Devlin

Carl Friedrich Gauss: Titan of Science, by G. Waldo Dunnington, with additional material by Jeremy Gray and Fritz-Egbert Dohse

The Changing Space of Geometry, edited by Chris Pritchard

Circles: A Mathematical View, by Dan Pedoe

Complex Numbers and Geometry, by Liang-shin Hahn

Cryptology, by Albrecht Beutelspacher

Five Hundred Mathematical Challenges, Edward J. Barbeau, Murray S. Klamkin, and William O. J. Moser

From Zero to Infinity, by Constance Reid

The Golden Section, by Hans Walser. Translated from the original German by Peter Hilton, with the assistance of Jean Pedersen.

I Want to Be a Mathematician, by Paul R. Halmos

Journey into Geometries, by Marta Sved

JULIA: a life in mathematics, by Constance Reid

The Lighter Side of Mathematics: Proceedings of the Eugène Strens Memorial Conference on Recreational Mathematics & Its History, edited by Richard K. Guy and Robert E. Woodrow

Lure of the Integers, by Joe Roberts

Magic Tricks, Card Shuffling, and Dynamic Computer Memories: The Mathematics of the Perfect Shuffle, by S. Brent Morris

The Math Chat Book, by Frank Morgan

Mathematical Apocrypha, by Steven G. Krantz

Mathematical Carnival, by Martin Gardner

Mathematical Circles Vol I: In Mathematical Circles Quadrants I, II, III, IV, by Howard W. Eves

Mathematical Circles Vol II: Mathematical Circles Revisited and Mathematical Circles Squared, by Howard W. Eves

Mathematical Circles Vol III: Mathematical Circles Adieu and Return to Mathematical Circles, by Howard W. Eves

Mathematical Circus, by Martin Gardner

Mathematical Cranks, by Underwood Dudley

Mathematical Evolutions, edited by Abe Shenitzer and John Stillwell

Mathematical Fallacies, Flaws, and Flimflam, by Edward J. Barbeau

Mathematical Magic Show, by Martin Gardner

Mathematical Reminiscences, by Howard Eves

Mathematical Treks: From Surreal Numbers to Magic Circles, by Ivars Peterson

Mathematics: Queen and Servant of Science, by E.T. Bell

Memorabilia Mathematica, by Robert Edouard Moritz

New Mathematical Diversions, by Martin Gardner

Non-Euclidean Geometry, by H. S. M. Coxeter

Numerical Methods That Work, by Forman Acton
Numerology or What Pythagoras Wrought, by Underwood Dudley
Out of the Mouths of Mathematicians, by Rosemary Schmalz
Penrose Tiles to Trapdoor Ciphers . . . and the Return of Dr. Matrix, by Martin Gardner
Polyominoes, by George Martin
Power Play, by Edward J. Barbeau
The Random Walks of George Pólya, by Gerald L. Alexanderson
Remarkable Mathematicians, from Euler to von Neumann, Ioan James
The Search for E.T. Bell, also known as John Taine, by Constance Reid
Shaping Space, edited by Marjorie Senechal and George Fleck
Sherlock Holmes in Babylon and Other Tales of Mathematical History, edited by Marlow Anderson, Victor Katz, and Robin Wilson
Student Research Projects in Calculus, by Marcus Cohen, Arthur Knoebel, Edward D. Gaughan, Douglas S. Kurtz, and David Pengelley
Symmetry, by Hans Walser. Translated from the original German by Peter Hilton, with the assistance of Jean Pedersen.
The Trisectors, by Underwood Dudley
Twenty Years Before the Blackboard, by Michael Stueben with Diane Sandford
The Words of Mathematics, by Steven Schwartzman

MAA Service Center
P.O. Box 91112
Washington, DC 20090-1112
800-331-1622 FAX 301-206-9789

Introduction

For the past one hundred years, the Mathematical Association of America has been publishing high-quality articles on the history of mathematics, some written by distinguished historians such as Florian Cajori, Julian Lowell Coolidge, Max Dehn, David Eugene Smith, Carl Boyer, and others. Many well-known historians of the present day also contribute to the MAA's journals. Some years ago, Robin Wilson and Marlow Anderson, along with the late John Fauvel, a distinguished and sorely missed historian of mathematics, decided that it would be useful to reprint a selection of these papers and to set them in the context of modern historical research, so that current mathematicians can continue to enjoy them and so that newer articles can be easily compared with older ones. After John's untimely death, Victor Katz was asked to fill in and help bring this project to completion.

A careful reading of some of the older papers in particular shows that although modern research has introduced some new information or has fostered some new interpretations, in large measure they are neither dated nor obsolete. Nevertheless, we have sometimes decided to include two or more papers on a single topic, written years apart, to show the progress in the history of mathematics.

The editors hope that you will enjoy this collection covering nearly four thousand years of history, from ancient Babylonia up to the time of Euler in the eighteenth century. We wish to thank Don Albers, Director of Publication at the MAA, and Gerald Alexanderson, chair of the publications committee of the MAA, for their support for the history of mathematics at the MAA in general, and for this project in particular. We also want to thank Beverly Ruedi for her technical expertise in preparing this volume for publication.

Contents

Introduction	vii
--------------	-----

Ancient Mathematics

Foreword	3
Sherlock Holmes in Babylon, <i>R. Creighton Buck</i>	5
Words and Pictures: New Light on Plimpton 322, <i>Eleanor Robson</i>	14
Mathematics, 600 B.C.–600 A.D., <i>Max Dehn</i>	27
Diophantus of Alexandria, <i>J. D. Swift</i>	41
Hypatia of Alexandria, <i>A. W. Richeson</i>	47
Hypatia and Her Mathematics, <i>Michael A. B. Deakin</i>	52
The Evolution of Mathematics in Ancient China, <i>Frank Swetz</i>	60
Liu Hui and the First Golden Age of Chinese Mathematics, <i>Philip D. Straffin, Jr.</i>	69
Number Systems of the North American Indians, <i>W. C. Eells</i>	83
The Number System of the Mayas, <i>A. W. Richeson</i>	94
Before The Conquest, <i>Marcia Ascher</i>	98
Afterword	105

Medieval and Renaissance Mathematics

Foreword	109
The Discovery of the Series Formula for π by Leibniz, Gregory and Nilakantha, <i>Ranjan Roy</i>	111
Ideas of Calculus in Islam and India, <i>Victor J. Katz</i>	122
Was Calculus Invented in India?, <i>David Bressoud</i>	131
An Early Iterative Method for the Determination of $\sin 1^\circ$, <i>Farhad Riahi</i>	138
Leonardo of Pisa and his Liber Quadratorum, <i>R. B. McClenon</i>	143
The Algorists vs. the Abacists: An Ancient Controversy on the Use of Calculators, <i>Barbara E. Reynolds</i>	148
Sidelights on the Cardan-Tartaglia Controversy, <i>Martin A. Nordgaard</i>	153
Reading Bombelli's x -purgated Algebra, <i>Abraham Arcavi and Maxim Bruckheimer</i>	164
The First Work on Mathematics Printed in the New World, <i>David Eugene Smith</i>	169
Afterword	173

The Seventeenth Century

Foreword	177
An Application of Geography to Mathematics: History of the Integral of the Secant, <i>V. Frederick Rickey and Philip M. Tuchinsky</i>	179
Some Historical Notes on the Cycloid, <i>E. A. Whitman</i>	183
Descartes and Problem-Solving, <i>Judith Grabiner</i>	188

René Descartes' Curve-Drawing Devices: Experiments in the Relations Between Mechanical Motion and Symbolic Language, <i>David Dennis</i>	199
Certain Mathematical Achievements of James Gregory, <i>Max Dehn and E. D. Hellinger</i>	208
The Changing Concept of Change: The Derivative from Fermat to Weierstrass, <i>Judith V. Grabiner</i>	218
The Crooked Made Straight: Roberval and Newton on Tangents, <i>Paul R. Wolfson</i>	228
On the Discovery of the Logarithmic Series and Its Development in England up to Cotes, <i>Josef Ehrenfried Hofmann</i>	235
Isaac Newton: Man, Myth, and Mathematics, <i>V. Frederick Rickey</i>	240
Reading the Master: Newton and the Birth of Celestial Mechanics, <i>Bruce Pourciau</i>	261
Newton as an Originator of Polar Coordinates, <i>C. B. Boyer</i>	274
Newton's Method for Resolving Affected Equations, <i>Chris Christensen</i>	279
A Contribution of Leibniz to the History of Complex Numbers, <i>R. B. McClenon</i>	288
Functions of a Curve: Leibniz's Original Notion of Functions and Its Meaning for the Parabola, <i>David Dennis and Jere Confrey</i>	292
Afterword	297

The Eighteenth Century

Foreword	301
Brook Taylor and the Mathematical Theory of Linear Perspective, <i>P. S. Jones</i>	303
Was Newton's Calculus a Dead End? The Continental Influence of Maclaurin's Treatise of Fluxions, <i>Judith Grabiner</i>	310
Discussion of Fluxions: from Berkeley to Woodhouse, <i>Florian Cajori</i>	325
The Bernoullis and the Harmonic Series, <i>William Dunham</i>	332
Leonhard Euler 1707–1783, <i>J. J. Burckhardt</i>	336
The Number e , <i>J. L. Coolidge</i>	346
Euler's Vision of a General Partial Differential Calculus for a Generalized Kind of Function, <i>Jesper Lützen</i>	354
Euler and the Fundamental Theorem of Algebra, <i>William Dunham</i>	361
Euler and Differentials, <i>Anthony P. Ferzola</i>	369
Euler and Quadratic Reciprocity, <i>Harold M. Edwards</i>	375
Afterword	383
Index	385
About the Editors	387



**Ancient
Mathematics**

Foreword

The twentieth century saw great strides in our understanding of the mathematics of ancient times. This was often achieved through the combined work of archaeologists, philologists, and historians of mathematics.

We especially see how this understanding has grown in the study of the mathematics of Mesopotamia. Although the clay tablets on which this mathematics was written were excavated beginning in the nineteenth century, it was not until early in the twentieth century that a careful study of the mathematics on some of these tablets was undertaken. In particular, the tablet known as Plimpton 322 was first published by Neugebauer and Sachs in 1945, who determined that the numbers in each row of the tablet always included two out of the three numbers of a Pythagorean triple. Since that time, there has been a great scholarly debate on how those numbers were found, as well as the general purpose of the tablet. In our opening papers, we present two discussions of this issue, one by R. Creighton Buck and a second by Eleanor Robson. Both of these papers illustrate the necessity of applying ideas from several disciplines to help us make sense of the past.

Greek mathematics has, of course, been studied ever since the demise of Greek civilization. A survey of the history of Greek mathematics, as it was understood in the 1940s, is presented here by Max Dehn, a prominent mathematician in his own right—he solved one of the Hilbert problems. Dehn’s article originally appeared in four parts in the *Monthly*, each part dealing with a different chronological period. The first part considers the work of Pythagoras and his school; the second deals with Euclid; the third considers Apollonius and Archimedes; while the fourth gives us a summary of the mathematics in Greek culture under the domination of the Roman empire.

Two of the mathematicians mentioned by Dehn are dealt with in more detail in the following articles, one on Diophantus and two on Hypatia. J. D. Swift examines several problems posed by Diophantus and explains some of his ingenious solutions. A. W. Richeson discusses the life of Hypatia through a detailed analysis of the sources available to him in 1940. In a more recent article, Michael Deakin considers the latest research on the work of Hypatia. He explains how we know what we do know, especially in regard to her mathematical work, and what remains as speculation.

Frank Swetz presents a detailed survey of what we know about mathematics in ancient China. Not only does he explain in detail certain mathematical techniques of the Chinese, but he also presents a detailed bibliography so that the reader may explore further. Swetz briefly mentions the third-century mathematician Liu Hui, whose work is explored in greater detail by Philip Straffin in the next article. There we learn not only about Liu Hui’s commentaries and extensions of the Chinese classic *Nine Chapters on the Mathematical Art*, but also about Liu’s use of a limit argument to determine the volume of a pyramid. Although Liu used what is now called Cavalieri’s principle to determine certain volumes, he could not figure out how to determine the volume of

a sphere. Straffin shows us how a later mathematician, Zu Gengzhi, ultimately determined the correct formula for that volume through a creative use of the same principle.

This section concludes with three discussions of mathematics in the Americas. First, W. C. Eells reports on the data from many years of linguistic study and analyzes the structure of the number systems in numerous groups of North American Indians. Next, A. W. Richeson looks at the number system of Mayas, displaying both the head-variant form of the monuments as well as the more familiar written form in the codices. Marcia Ascher, in an article written to commemorate the 500th anniversary of Columbus's first visit to the western hemisphere, then discusses the mathematics of two of the civilizations living there at the time. She explains the quipus of the Incas of Peru and Ecuador and then deals anew with the Mayans, concentrating in particular on the types of mathematical problems that they could solve in their number system.

Sherlock Holmes in Babylon

R. CREIGHTON BUCK

American Mathematical Monthly 87 (1980), 335–345

Let me begin by clarifying the title “Sherlock Holmes in Babylon.” Lest some members of the Baker Street Irregulars be misled, my topic is the archaeology of mathematics, and my objective is to retrace a small portion of the research of two scholars: Otto Neugebauer, who is a recipient of the Distinguished Service Award, given to him by the Mathematical Association of America in 1979, and his colleague and long-time collaborator, Abraham Sachs. It is also a chance for me to repay both of them a personal debt. I went to Brown University in 1947, and as a new Assistant Professor I was welcomed as a regular visitor to the Seminar in the History of Mathematics and Astronomy. There, with a handful of others, I was privileged to watch experts engaged in the intellectual challenge of reconstructing pieces of a culture from random fragments of the past. (See [4], [5].)

This experience left its mark upon me. While I do not regard myself as a historian in any sense, I have always remained a “friend of the history of mathematics”; and it is in this role that I come to you today. Let me begin with a sample of the raw materials. Figure 1 is a copy of a cuneiform tablet measuring perhaps 3 inches by 5. The markings can be made by pressing the end of a cut reed into wet clay. Dating such a tablet is seldom easy. The appearance of this tablet suggests that it may have been made in Akkad in the city of Nippur in the year -1700 , about 3700 years ago.

Confronted with an artifact from an ancient culture, one asks several questions:

- (i) What is this and what are its properties?
- (ii) What was its original purpose?
- (iii) What does this tell me about the culture that produced it?

In the History of Science, one expects neither theorems nor rigorous proofs. The subject is replete with conjectures and even speculations; and in place of proof, one often finds mere confirmation: “I believe P implies Q ; and because I also believe Q , I therefore also believe P .”

In Figure 1, we draw a vertical line to separate the first two columns. In the first column, we recognize what seem to be counting symbols for the numbers from 1 through 9. Paired with these in the second column we see 9, then 1 and 8, then 2 and 7, and then 3 and 6. This suggests that what we have is a “table of 9’s”, a multiplication table for the factor 9. Checking further, we see 5 and 4 across from the

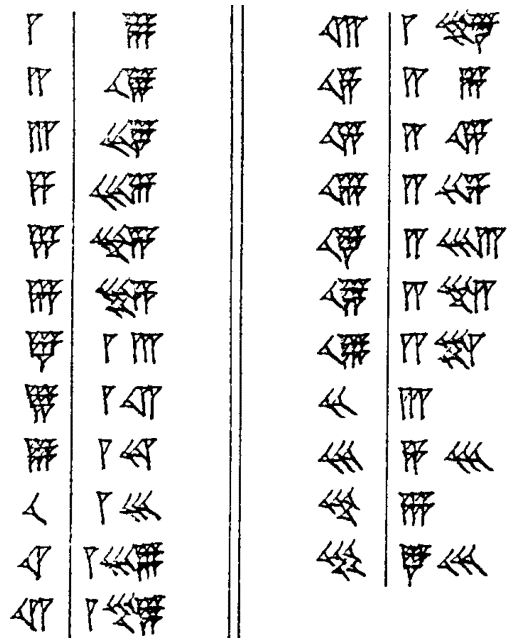


Figure 1.

counting symbol for 6, which confirms the conjecture. However, in the next line we see 7 and then across from it what seems to be a 1 and a 3.

We modify our conjecture; instead of an ordinary decimal system, we are dealing with a hybrid. There is a decimal substratum, using one type of wedge for units and another for tens but the system is base 60 in the large. The 1 and 3 in fact represent $60 + 3 = 63$. We then immediately conjecture that the same wedge symbol will be used for 1, for 60, for $(60)^2$, $(60)^3$, and so on, while the digits will be given in a decimal form.

Thus from a single tablet we might have conjectured a complete sexagesimal numeral system. We would then seek confirmation of this by examining other tablets, hoping to see the same patterns there. Indeed, this was done in the last century, and among the thousands of Babylonian tablets many were found that bear multiplication tables of the same general type as that given in Figure 1, generated by various multiplication factors. There are a great many duplicates.

We find the Babylonian numeral system cumbersome to write. In this paper, base 60 numerals will be written by putting the digits (0 through 59) in ordinary Arabic base ten, separating consecutive digits by the symbol “/”. The “units place” will be on the right as usual. Thus

$$7/13/28 \text{ represents } 28 + 13(60) + 7(60)^2 = 26,008.$$

Addition is easy:

$$\begin{array}{r} 14/28/31 \\ 3/35/45 \\ \hline 18/4/16 \end{array}$$

If the tablets that bear multiplication tables are catalogued, something strange is seen. Many tables of 9’s, 12’s, etc., are found; but there are also multiplication tables for unlikely factors, while many tables we would have expected never appear. In Figure 2, we list those that occur frequently.

We are left with three puzzles:

- (i) Why are some tables missing? (For example, 7, 11, 13, 14, etc.?)
- (ii) Why are there tables with factors such as $3/45$, $7/12$, $7/30$, and $44/26/40$?
- (iii) Why are there so many tablets with exactly the same multiplication tables on them?

Some clues are found; for example, there are tablets that contain two versions of the same multiplication

2	18	$1/15 = 75$	$7/12 = 432$
3	20	$1/20 = 80$	$7/30 = 450$
4	24	$1/30 = 90$	$8/20 = 500$
5	25	$1/40 = 100$	$12/30 = 750$
6	30	$2/15 = 135$	$16/40 = 1000$
8	36	$2/24 = 144$	$22/30 = 1350$
9	40	$2/30 = 150$	$44/26/40 = 160,000$
10	45	$3/20 = 200$	
12	48	$3/45 = 225$	and a scattering of others
15	50	$4/30 = 270$	
16		$6/40 = 400$	

Figure 2. Factors Used for Multiplication Tables

table, one done neatly and one less neatly and perhaps with an error or two. I am sure that a familiar picture comes immediately to your mind: a cluster of students, all engaged in copying a model table provided by the teacher who will shortly be grading their efforts. Are we not correct to infer that in Nippur there was probably an extensive school for scribes who were in training to become bureaucrats or priests?

To help answer the first two questions, let us examine another tablet, which for convenience I have transcribed into the slash notation. (See Figure 3.) This again fits the pattern of two matched columns, and we look for an explanation. We note at once that in the first few rows the product of the adjacent column numbers is always 60. There seem to be some exceptions, however. With the pair 9 and $6/40$, this product is

$$(9) \times (6/40) = (9) \times (400) = 3600$$

2	30	16	$3/45$	45	$1/20$
3	20	18	$3/20$	48	$1/15$
4	15	20	3	50	$1/12$
5	12	24	$2/30$	54	$1/6/40$
6	10	25	$2/24$	$1/4$	$56/15$
8	$7/30$	27	$2/13/20$	$1/12$	50
9	$6/40$	30	2	$1/15$	48
10	6	32	$1/52/30$	$1/20$	45
12	5	36	$1/40$	$1/21$	$44/26/40$
15	4	40	$1/30$		

Figure 3.

and again

$$(16) \times (3/45) = (16) \times (225) = 3600$$

while still further down, we see

$$(27) \times (2/13/20) = (27) \times (8000) = 216,000.$$

The solution becomes obvious if we write these products in Babylonian form; since 60 is 1/0, 3600 is 1/0/0, and 216,000 is 1/0/0/0. For confirmation, look at the last entry in the table:

$$\begin{aligned} (1/21) \times (44/26/40) &= (81) \times (160,000) \\ &= 12,960,000 \\ &= 1/0/0/0/0. \end{aligned}$$

If we now follow the Babylonian practice of omitting terminal zeros, we see that Figure 3 is merely a table of reciprocals, written in “sexagesimal floating point.” If A is an integer in the first column, the integer paired with it in the second column, A^R , is one chosen so that their product would be written as “1,” meaning *any* suitable power of 60. The integers that appear in the table will always be factorable into powers of 2, 3, and 5, since these have terminating reciprocals in base 60. The term “floating-point arithmetic” is today a computer concept but is also understandable to anyone who has used a slide rule or worked with logarithms; the concept would also have been familiar to medieval astronomers who multiplied large numbers by the device called “prosthaphaeresis.”

Now that Figure 3 is understood, we can answer the two puzzles left hanging on the previous page. Observe that the integers used to generate multiplication tables, as seen in Figure 2, mostly come from the standard reciprocal table. (There are also tablets that contain nonstandard reciprocals, reciprocals of such numbers as 7, 11, etc., of necessity given in terminating approximate form.) In floating point, $B \div A = B \times A^R$. Thus the combination of a set of multiplication tables and a reciprocal table makes it easy to carry out floating-point division, provided that the divisor is one of the “nice” numbers in base 60, of the form $2^\alpha 3^\beta 5^\gamma$. For example, let us divide 417 by 24; in base 60, this will be $6/57 \div 24 = 17/22/30$.

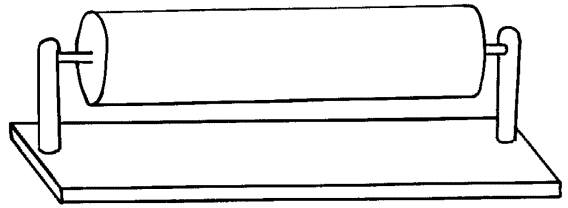
Method:

$$6/57 \div 24 = (6/57) \times (24)^R = (6/57) \times (2/30) :$$

$$6/57 \times 2 = 12 + 1/54 = 13/54$$

$$6/57 \times 30 = 3 + 28/30 = \underline{3/28/30}$$

$$\text{answer} = 17/22/30$$



2	1/40	3	1/30
3/45	3/20	4	2/30
5	4/30	4	3/45
7/12	6/40	6	
8/20	8	7/30	
12	10	9	
16	15	12/30	
20	18	16/40	
25	24	22/30	
40	36	30	
48	45	44/26/40	
Reciprocals	50	48	

Figure 4.

The last steps in this calculation are easier if one recalls that $30 = 2^R$, so that multiplication by 30 is the same as halving. (Of course the scribe must be sure to keep track of the actual magnitudes and place values.)

That common calculations were made in this fashion becomes even more plausible in the light of one remarkable discovery. This is an inscribed cylinder, carrying on its curved face a copy of the standard reciprocal table and each of the standard multiplication tables. (In Figure 4, we show this restored, with each multiplication table indicated by its generator.) With the help of this cylinder, perhaps mounted on a stand, a scribe could easily keep track of taxes and calculate wages; perhaps we have here the Babylonian version of a slide rule or desk calculator!

With this brief introduction to the arithmetic of the Babylonians, we turn to another tablet whose mathematical nature had been overlooked until the work of Neugebauer and Sachs. It is in the George A. Plimpton Collection, Rare Book and Manuscript Library, at Columbia University, and usually called Plimpton 322. (See Figure 5, which is reproduced here by permission of the Library.) The left side of this tablet has some erosion; traces of modern glue on the left edge suggest that a portion that had originally been attached there has since been lost or stolen. Since it was bought in a marketplace, one may only conjecture about its true origin and date, although the style suggests about -1600 for the latter. As with most such tablets, this had been assumed to be a commercial account or inventory report. We will attempt to show why one can be led to believe otherwise.

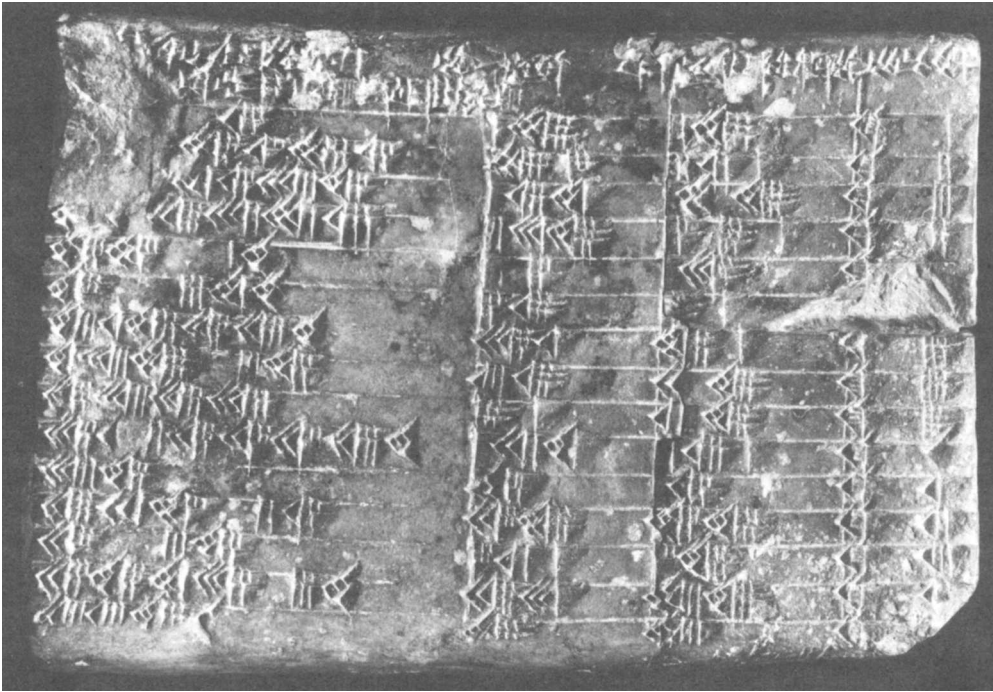


Figure 5. Plimpton 322

Column A	Column B	Column C
15	1/59	2/49
58/14/50/6/15	56/7	3/12/1
1/15/33/45	1/16/41	1/50/49
5 29/32/52/16	3/31/49	5/9/1
48/54/ 1/40	1/5	1/37
47/ 6/41/40	5/19	8/1
43/11/56/28/26/40	38/11	59/1
41/33/59/ 3/45	13/19	20/49
38/33/36/36	9/1	12/49
35/10/2/28/27/24/26/40	1/22/41	2/16/1
33/45	45	1/15
29/21/54/ 2/15	27/9	48/49
27/ 3/45	7/12/1	4/49
25/48/51/35/6/40	29/31	53/49
23/13/46/40	56	53

Figure 6. Plimpton 322

First, let us transcribe it into the slash notation, as seen in Figure 6. We have reproduced the three main columns, which we have labeled *A*, *B*, and *C*. We note that there are gaps in column *A*, due to the erosion. However, it seems apparent that the numbers there are steadily decreasing. We note that

some of the numerals there are short and some long, apparently at random. In contrast with this, all the numerals in columns *B* and *C* are rather short, and we do not see any evidence of general monotonicity.

Since it is easier for us to work with Arabic numerals, let us translate columns *B* and *C* into these numerals and look for patterns. (See Figure 7.) We see at once that *B* is smaller than *C*, with only two exceptions. Also, playing with these numbers, we find that column *B* contains exactly one prime, namely, 541, while column *C* contains eight numbers that are prime.

In the first 20,000 integers, there are about 2,300 primes, which is about 10 percent; among 15 inte-

<i>B</i>	<i>C</i>	<i>B</i>	<i>C</i>
119	169	541	769
3367	11521	4961	8161
4601	6649	45	75
12709	18541	1679	2929
65	97	25921	289
319	481	1771	3229
2291	3541	56	53
799	1249		

Figure 7.

$C + B$	$C - B$
288	50
14888	8154
11250	2048
31250	5832
162	32
800	162
5832	1250
2048	450
1310	228
13132	3200
120	30
4608	1250
26210	-25632
5000	1458
109	-3

Figure 8.

B	C	(a, b)
119	169	12,5
3367	11521	?
4601	6649	75,32
12709	18541	125,54
65	97	9, 4
319	481	20,9
2291	3541	54,25
799	1249	32,15
541	769	?
4961	8161	81,40
45	75	?
1679	2929	48,25
25921	289	?
1771	3229	50,27
56	53	?

Figure 9.

gers, selected at random from this interval, we might, then, expect to see one or two primes, but certainly not eight! This at once tells us that the tablet is mathematical and not merely arithmetical. (Imagine your feelings if you were to find a Babylonian tablet with a list of the orders of the first few sporadic simple groups.)

Encouraged, one attempts to find further visible patterns, for example, by combining the entries in columns B and C in various ways. One of the earliest tries is immediately successful. In Figure 8, we show the results of calculating $C + B$ and $C - B$. If you are sensitive to arithmetic you will note that, in almost every case, the numbers are each twice a perfect square.

If $C + B = 2a^2$ and $C - B = 2b^2$, then $B = a^2 - b^2$ and $C = a^2 + b^2$. Thus the entries in these columns could have been generated from integer pairs (a, b) . In passing, we note that B , being $(a-b)(a+b)$, is not apt to be prime; on the other hand, when a and b are relatively prime, every prime of the form $4N + 1$ can be expressed as $a^2 + b^2$.

In Figure 9, we have recopied columns B and C , together with the appropriate pairs (a, b) in the cases where this representation is possible. As a further confirmation that we are on the right track, we note that in every such pair the numbers a and b are both “nice”, that is, factorable in terms of 2, 3, and 5. In five cases, the pattern breaks down and no pair exists. It will be a further confirmation if we can explain these discrepancies as errors made by the scribe who produced the tablet. We make a simple hypothesis and assume that B and C were each

computed independently from the pair (a, b) and that a few errors were made but each affected only one number in each row. Thus in each vacant place we will assume that either B or C is correct and the other wrong, and attempt to restore the correct entry. Since we do not know the correct pair (a, b) we must find it; because of the evidence in the rest of the table, we insist that an acceptable pair must be composed of “nice” sexagesimals.

We start with line 9; here, $B = 541$, which happens to be the only prime in Column B . We therefore assume B is wrong and C is correct, and thus write $C = 769 = a^2 + b^2$. This has a single solution, the pair $(25, 12)$. (We also note that both happen to be nice sexagesimals.) If this is correct, then B should have been $(25)^2 - (12)^2 = 481$, instead of 541 as given. Is there an obvious explanation for this mistake? Yes, for in slash notation, $541 = 9/1$ and $481 = 8/1$. The anomaly in line 9 seems to be merely a copy error.

Turn now to line 13; here, B is far larger than C , which is contrary to the pattern. Assume that B is in error and C is correct, and again try $C = 289 = a^2 + b^2$. There is a “nice” unique solution, $(15, 8)$, and using these, we are led to conjecture that the correct value of B is $(15)^2 - (8)^2 = 161$. Again, we ask if there is an obvious explanation for arriving at the incorrect value given, 25921. A partial answer is immediate: $(161)^2 = 25921$; so that for some reason the scribe recorded the *square* of the correct value for B .

Continuing, consider line 15. Since $B = 56$ and $C = 53$, we have $B > C$, which does not match the

general pattern. However, it is not clear whether B is too large or C too small. Trying the first, we assume C is correct and solve $53 = a^2 + b^2$, obtaining the unique answer $(7, 2)$. We reject this, since 7 is not a nice sexagesimal. Now assume that B is correct, and write $56 = a^2 - b^2 = (a + b)(a - b)$. This has two solutions, $(15, 13)$ and $(9, 5)$. We reject the first and use the second, obtaining $92 + 52 = 106$ as the correct value of C . Seeking an explanation, we note that the value given by the scribe, 53, is exactly half of the correct value.

Turning now to line 2 of Figure 9, we have $B = 3367$ and $C = 11521$, either of which might be correct. Assume that $C = a^2 + b^2$ and find two solutions $(100, 39)$ and $(89, 60)$. While 100 and 60 are nice, 39 and 89 are not, so we reject both pairs and assume that B is correct. Writing $3367 = (a - b)(a + b)$ and factoring 3367 in all ways, we find four pairs: $(1684, 1683)$, $(244, 237)$, $(136, 123)$, $(64, 27)$, of which we can accept only the last. This yields $(64)^2 + (27)^2 = 4825$ as the correct C . Comparing this with the number 11521 that appeared on the tablet, we see no immediate naive explanation for the error. For example, since $4825 = 1/20/25$ and $11521 = 3/12/1$, it does not seem to be a copy error. Without an explanation, we may have a little less confidence in this reconstruction of the entries in line 2.

The last misfit in the table is line 11, where we have $B = 45$ and $C = 75$. This is unusual also because this is the only case where B and C have a common factor. The sums-and-differences-of-squares pattern failed because neither $C + B = 120$ nor $C - B = 30$ is twice a square. However, everything becomes clearer if we go back to base 60 notation and remember that we use floating point; for $120 = 2/0$, which is twice $1/0$ and which we can also write as 1 , clearly a perfect square. In the same way, 30 is twice 15 , which is also 4^R and which is the square of 2^R . The pattern is preserved and no corrections need be made in the entries: with $a = 1 = 1/0$ and $b = \frac{1}{2} = 2^R = 30 = 0/30$, we have $a^2 = 1/0$ and $b^2 = 0/15$, and

$$C = a^2 + b^2 = 1/0 + 0/15 = 1/15 = 75$$

$$B = a^2 - b^2 = 1/0 - 0/15 = 0/45 = 45.$$

(Another aspect of the line 11 entries will appear later.)

With this, we have completed the work of editing the original tablet. In Figure 10, we give a corrected table for columns B and C , together with the appropriate pairs (a, b) from which they can be calculated.

B	C	(a, b)
119	169	12, 5
3367	4825	64, 27
4601	6649	75, 32
12709	1854	125, 54
65	97	9, 4
319	481	20, 9
2291	3541	54, 25
799	1249	32, 15
481	769	25, 12
4961	8161	81, 40
45	75	$1, \frac{1}{2} = 30$
1679	2929	48, 25
161	289	15, 8
1771	3229	50, 27
56	106	9, 5

Figure 10. Corrected Version

It is now the time to raise the second canonical question: What was the purpose behind this tablet? Speculation in this direction is less restricted, since the road is not as well marked. We can begin by asking if numbers of the form $a^2 - b^2$ and $a^2 + b^2$ have any special properties. In doing so, we run the risk of looking at ancient Babylonia from the twentieth century, rather than trying to adopt an autochthonous viewpoint. Nevertheless, one relation is extremely suggestive, involving both algebra and geometry. For any numbers (integers) a and b ,

$$(a^2 - b^2)^2 + (2ab)^2 = (a^2 + b^2)^2. \tag{1}$$

In addition, if we introduce $D = 2ab$, then B , C , and D can form a right-angled triangle with $B^2 + D^2 = C^2$. And finally, these formulas generate all Pythagorean triplets (triangles) from the integer parameters (a, b) . (See Figure 11.)

There is no independent information showing that these facts were known to the Babylonians at the time we conjecture that this tablet was inscribed, although, as will appear later, their algebra had already

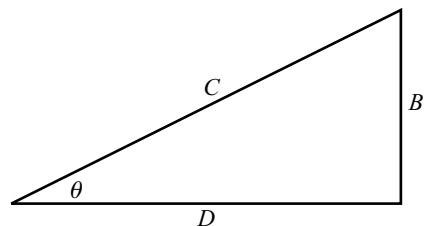


Figure 11. $B = a^2 - b^2, D = 2ab, C = a^2 + b^2$

mastered the solution of quadratic equations. If the tablet indeed is connected with this observation, then the unknown column A numbers ought to be connected in some way with the same triangle. The next step is, then, to proceed as before and try many different combinations of B , C , and D , in hopes that one of these will approximate the entries in column A . Slopes and ratios are an obvious starting point, so one calculates $C \div B$, $C \div D$, $B \div D$, etc. After discarding many failures, one arrives at the combination $(B \div D)^2$. In Figure 12, we give the values of this expression, calculated from the corrected values of B and using the hypothetical values of (a, b) to find D . (We remark that it was very helpful to have a programmable pocket calculator that could be trained to work in sexagesimal arithmetic!)

If we now return to Figure 6 and compare the numerals given there in column A with those that appear in Figure 12, we see that there is almost total agreement. For example, in line 10 we have exact duplication of an eight-digit sexagesimal! On probabilistic grounds alone, this is an overwhelming confirmation. Of course, at the top of the tablet where there were gaps due to erosion, Figures 6 and 12 are not the same, but it is evident that the calculated data in Figure 12 can be regarded as filling in the gaps. There are two minor disagreements in the two tables. In line 13, the tablet does not show an internal “0” that is present in Figure 12. This could have been the custom of the scribe in dealing with such

line	value
1	59/0/15
2	56/56/58/14/50/6/15
3	55/7/41/15/33/45
4	53/10/29/32/52/16
5	48/54/1/40
6	47/6/41/40
7	43/11/56/28/26/40
8	41/33/45/14/3/45
9	38/33/36/36
10	35/10/2/28/27/24/26/40
11	33/45
12	29/21/54/2/15
13	27/0/3/45
14	25/48/51/35/6/40
15	23/13/46/40

Figure 12. Calculated Values of $(B \div D)^2$

an event. In line 8, the scribe has written a digit “59” where there should have been a consecutive pair of digits, “45/14”. Since $59 = 45 + 14$, it is not difficult to invent several different ways in which an error of this sort could have been made.

It should be remarked that Neugebauer and Sachs did not use $(B \div D)^2$ as a source for column A but rather $(C \div D)^2$. Because of the relationship between B and C , and formula (1), one sees that $(C \div D)^2 = (B \div D)^2 + 1$. Thus, the only effect of the change would be to introduce an initial “1/” before all the sexagesimals that appear in Figure 12, and the reason for their choice was that they believed that this was true for column A on the Plimpton tablet. Others who have examined the tablet do not agree. (I have not seen the tablet, and I do not believe it matters which alternative is used.)

We now know the relationship of columns A , B , and C . Referring to Figure 11, C is the hypotenuse, B the vertical side, and A is the square of the slope of the triangle; thus, in modern notation $A = \tan^2 \theta$. It is interesting to observe that the anomalous case of line 11, with $B = 45$ and $C = 75$, turns out to be the familiar 3, 4, 5 triangle; in the Babylonian case, this would seem to have been the $\frac{3}{4}, 1, \frac{5}{4}$ triangle, since $45 = 3 \times 4^R$ and $75 = 1/15 = 5 \times 4^R$. Of course the triangle, the side D , and the parameters (a, b) are all constructs of ours and not immediately visible in the original tablet. All that we can assert without controversy is that $A = B^2 \div (C^2 - B^2)$.

Let us reexamine some of our reasoning. In lines 2, 9, 13, and 15, the scribe recorded correct values for A but incorrect values for C, B, B , and C , respectively. This suggests strongly that A was not calculated directly from the values of B and C , but that A, B , and C were all calculated independently from data that do not appear on the tablet; our hypothetical pair (a, b) gains life. (Of course there is the possibility that the tablet before us is merely a copy from another master tablet.) In either case, it seems odd that column A should be error free while columns B and C , involving simpler numbers, should have four errors.

Other questions can be raised. If, as argued by Neugebauer, the purpose of the tablet was to record a collection of integral-sided Pythagorean triangles (triplets), why do we not see the values of D , or at least the useful parameters (a, b) ? And why would one want the values in column A which are squares of the slope? And why should the entries be arranged in an order that makes the numbers A decrease monotonically?

Variants of this explanation have been proposed. If one computes the values of the angle θ for each line of the tablet, they are seen to decrease steadily from about 45° to about 30° , in steps of about 1° . Is this an accident? Could this tablet be a primitive trigonometric table, intended for engineering or astronomical use? But again, why is $\tan^2 \theta$ useful [3], [6]?

Additional confirmation of such a hypothesis could be given by an outline of a computational procedure leading to the tablet, which makes all of the errors plausible and also shows why they would have occurred preferentially in columns B and C . (See [1], [4], [7].)

Building upon an earlier suggestion of Bruins, an intriguing explanation has been recently proposed by Voils. In Nippur, a large number of “school texts” have been found, many containing arithmetic exercises. Among these, a standard puzzle problem is quite common. The student is given the difference (or sum) of an unknown number and its reciprocal and asked to find the number. If x is the number (called “igi”) and x^R is its reciprocal (called “igibi”), then the student is to solve the equation $x - x^R = d$. Thus, the “igi and igibi” problems are quadratic equations of a standard variety.

The school texts teach a specific solution algorithm: “Find half of d , square it, add 1, take the square root, and then add and subtract half of d .” This is easily seen to be nothing more than a version of the quadratic formula, tailored to the “igi and igibi” problems. Voils connects this class of problems, and the algorithm above, with the Plimpton tablet as follows.

First, assume with Bruins that the tablet was computed not from the pair (a, b) but from a single parameter, the number $x = a \div b$. Since a and b are both “nice”, the number x and its reciprocal x^R can each be calculated easily. Indeed, $x = a \times b^R$ and $x^R = b \times a^R$, and a^R and b^R each appear in a standard reciprocal table. Next observe that

$$\begin{aligned} B &= a^2 - b^2 = (ab)(x - x^R) \\ C &= a^2 + b^2 = (ab)(x + x^R) \\ A &= \left(\frac{B}{D}\right)^2 = \left\{\frac{1}{2}(x - x^R)\right\}^2. \end{aligned}$$

This shows that the entries A, B, C in the Plimpton tablet could have been easily calculated from a special reciprocal table that listed the paired values x and x^R . Indeed, the numbers B and C can be obtained from $x \pm x^R$ merely by multiplying these by

integers chosen to simplify the result and shorten the digit representation. (See [1], [2], [7].)

Voils adds to this suggestion of Bruins the observation that the numbers A are exactly the results obtained at the end of the second step in the solution algorithm, $(d/2)^2$, applied to an igi-igibi problem whose solution is x and x^R . Furthermore, the numbers B and C can be used to produce other problems of the same type but having the same intermediate results in the solution algorithm. Thus Voils proposes that the Plimpton tablet has nothing to do with Pythagorean triplets or trigonometry but, instead, is a pedagogical tool intended to help a mathematics teacher of the period make up a large number of igi-igibi quadratic equation exercises having known solutions and intermediate solution steps that are easily checked [7].

It is possible to point to another weak confirmation of this last approach. Suppose that we want a graduated table of numbers x and their reciprocals x^R . We start with the class of *all* pairs (a, b) of relatively prime integers such that $b < a < 100$ and each integer a and b is “nice”, factorable into powers of 2, 3, and 5. It is then easy to find the terminating Babylonian representation for both $x = a \div b$ and for $x^R = b \div a$. Make a table of these, arranged with x decreasing. Impose one further restriction:

$$\sqrt{3} < x < 1 + \sqrt{2}.$$

(This corresponds to the limitation $30^\circ < \theta < 45^\circ$, where θ is the base angle in the triangle in Figure 11.)

Then, the resulting list of pairs will coincide with that given in Figure 10, the corrected Plimpton table, except for three minor points. The pair $(16, 9)$ does not appear, the pair $(125, 54)$ does appear, and instead of the pair $(2, 1)$ we have the pair $(1, \frac{1}{2})$; in passing, we recall that the last pair yields the standard 3, 4, 5 Pythagorean triangle.

Unlike Doyle’s stories, this has no final resolution. Any of these reconstructions, if correct, throws light upon the degree of sophistication of the Babylonian mathematician and breathes life into what was otherwise dull arithmetic. For other vistas into the past, especially those that show us the beginnings of computational astronomy, I refer the reader to the bibliography. I can do no better than to close with an analogy used by Neugebauer:

In the “Cloisters” of the Metropolitan Museum in New York there hangs a magnificent tapestry which tells the tale of the Unicorn. At the end

we see the miraculous animal captured, gracefully resigned to his fate, standing in an enclosure surrounded by a neat little fence. This picture may serve as a simile for what we have attempted here. We have artfully erected from small bits of evidence the fence inside which we hope to have enclosed what may appear as a possible living creature. Reality, however, may be vastly different from the product of our imagination; perhaps it is vain to hope for anything more than a picture which is pleasing to the constructive mind, when we try to restore the past.

— *The Exact Sciences in Antiquity* (p. 177)

References

1. E. M. Bruins, On Plimpton 322: Pythagorean numbers in Babylonian mathematics, *Kon. Neder. Akad. U. Wetensch. Proceedings*, 52 (1949) 629–632.
2. E. M. Bruins, Pythagorean triads in Babylonian Mathematics, *Math. Gaz.*, 41 (1957) 25–28.
3. Howard Eves, *Introduction to the History of Mathematics*, 4th ed., Holt, Rinehart & Winston, New York, 1976.
4. O. Neugebauer, *The Exact Sciences in Antiquity*, Dover, New York, 1969.
5. O. Neugebauer and A. Sachs, *Mathematical Cuneiform Texts*, American Oriental Ser., vol. 29, American Oriental Society, New Haven, 1945.
6. Derek J. de S. Price, The Babylonian ‘Pythagorean triangle’ tablet, *Centaurus*, 10 (1964) 219–231.
7. D. L. Voils, to appear in *Historia Mathematica*.

Words and Pictures: New Light on Plimpton 322

ELEANOR ROBSON

American Mathematical Monthly 109 (2002), 105–120

1 Introduction

In this paper I shall discuss Plimpton 322, one of the world's most famous ancient mathematical artefacts [Figure 1]. But I also want to explore the ways in which studying ancient mathematics is, or should be, different from researching modern mathematics. One of the most cited analyses of Plimpton 322, published some twenty years ago, was called "Sherlock Holmes in Babylon" [4]. This enticing title gave out the message that deciphering historical documents was rather like solving a fictional murder

mystery: the amateur detective-historian need only pit his razor-sharp intellect against the clues provided by the self-contained story that is the piece of mathematics he is studying. Not only will he solve the puzzle, but he will outwit the well-meaning but incompetent professional history-police every time. In real life, the past isn't like an old-fashioned whodunnit: historical documents can *only* be understood in their historical context.

Let's start with a small experiment: ask a friend or colleague to draw a triangle. The chances are that he or she will draw an equilateral triangle with a

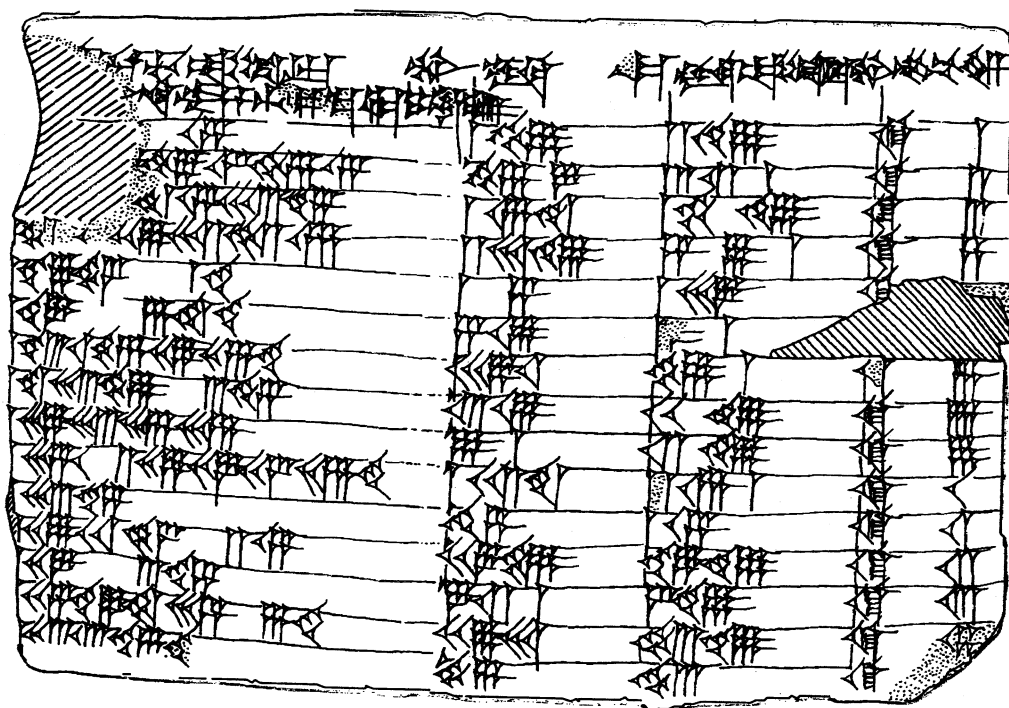


Figure 1. Plimpton 322 (obverse). Drawing by the author.

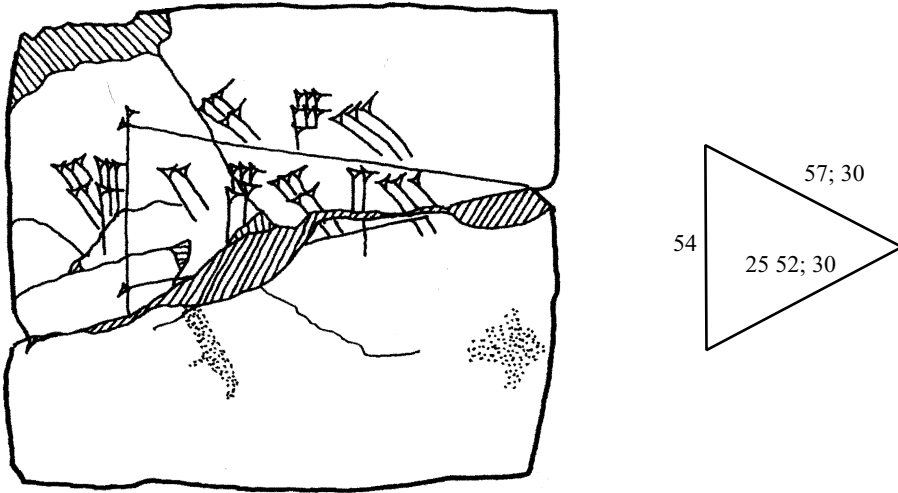


Figure 2. UM 29-15-709 (obverse). Drawing by the author [26, p. 29].

horizontal base. That is our culturally determined concept of an archetypal, perfect triangle. However, if we look at triangles drawn on ancient cuneiform tablets like Plimpton 322, we see that they all point right and are much longer than they are tall: very like a cuneiform wedge in fact. A typical example is UM 29-15-709, a scribal student's exercise, from ancient Nippur, in calculating the area of a triangle [Figure 2]. The scale drawing next to it shows how elongated the sketch is.

We tend to think of mathematics as relatively culture-free; i.e., as something that is *out there*, waiting to be discovered, rather than a set of socially agreed conventions. If a simple triangle can vary so much from culture to culture, though, what hope have we in relying on our modern mathematical sensibilities to interpret more complex ancient mathematics? Unlike Sherlock Holmes we cannot depend solely on our own intuitions and deductive powers, and we cannot interrogate the ancient authors or scrutinise their other writings for clues. We therefore have to exploit all possible available resources: language, history and archaeology, social context, as well as the network of mathematical concepts within which the artefact was created. In the case of Plimpton 322, for instance, there are three competing interpretations, all equally valid mathematically. As I shall show, it is these contextualising tools that enable us to choose between them.

Plimpton 322 is just one of several thousand mathematical documents surviving from ancient Iraq (also called Mesopotamia). In its current state, it comprises a four-column, fifteen-row table of

Pythagorean triples, written in cuneiform (wedge-shaped) script on a clay tablet measuring about 13 by 9 by 2 cm [20, Text A, pp. 38–41]. The handwriting of the headings is typical of documents from southern Iraq of 4000–3500 years ago. Its second and third columns list the smallest and largest member of each triple—we can think of them as the shortest side s and the hypotenuse d of a right-angled triangle—while the final column contains a line-count from 1 to 15. Part of the tablet has broken away at the beginning of the first column but, depending on whether you believe the column has fully survived or not, it holds the square of *either* the hypotenuse *or* the shortest side of the triangle divided by the square of the longer side l . Whether it lists d^2/l^2 or s^2/l^2 , this column is in descending numerical order. The numbers are written in the base 60, or sexagesimal, place value system. I shall transliterate them with a semicolon marking the boundary between integers and fractions, and spaces in between the other sexagesimal places [Figure 3].

There have been three major interpretations of the tablet's function since it was first published [Figure 4]:¹

1. Some have seen Plimpton 322 as a form of trigonometric table (e.g., [15]): if Columns II and

¹Incidentally, we can dismiss immediately any suspicion that Plimpton 322 might be connected with observational astronomy. Although some simple records of the movements of the moon and Venus *may* have been made for divination in the early second millennium BCE, the accurate and detailed programme of astronomical observations for which Mesopotamia is rightly famous began a thousand years later, at the court of the Assyrian kings in the eighth century BCE [3].

[<i>ta</i>]- <i>ki-il-ti ši-li-ip-tim</i>	İB.SI ₈ SAG	İB.SI ₈ <i>ši-li-ip-tim</i>	MU.BI.IM
[<i>sa</i> 1 <i>in</i>]- <i>na-as-sà-hu-ma SAG i-il-lu-ú</i>			
[(1) 59] 00 15	1 59	2 49	KI.1
[(1) 56 56] 58 14 50 06 15	56 07	1 20 25	KI.2
[(1) 55 07] 41 15 33 45	1 16 41	1 50 49	KI.3
(1) 53 10 29 32 52 16	3 31 49	5 09 01	KI.4
(1) 48 54 01 40	1 05	1 37	KI.[5]
(1) 47 06 41 40	5 19	8 01	[KI.6]
(1) 43 11 56 28 26 40	38 11	59 01	KI.7
(1) 41 33 45 14 3 45	13 19	20 49	KI.8
(1) 38 33 36 36	8 01	12 49	KI.9
(1) 35 10 02 28 27 24 26 40	1 22 41	2 16 01	KI.10
(1) 33 45	45	1 15	KI.11
(1) 29 21 54 2 15	27 59	48 49	KI.12
(1) 27 00 03 45	2 41	4 49	KI.13
(1) 25 48 51 35 6 40	29 31	53 49	KI.14
(1) 23 13 46 40	28	53	KI.15

Figure 3. Transliteration of Plimpton 322.

<i>line</i>	α	<i>p</i>	<i>q</i>	<i>x</i>	$1/x$
1	44.76°	12	5	2 24	25
2	44.25°	1 04	27	2 22 13 20	25 18 45
3	43.79°	1 15	32	2 20 37 30	25 36
4	43.27°	2 05	54	2 18 53 20	25 55 12
5	42.08°	9	4	2 15	26 40
6	41.54°	20	9	2 13 20	27
7	40.32°	54	25	2 09 36	27 46 40
8	39.77°	32	15	2 08	28 07 30
9	38.72°	25	12	2 05	28 48
10	37.44°	1 21	40	2 01 30	29 37 46 40
11	36.87°	2	1	2	30
12	34.98°	48	25	1 55 12	31 15
13	33.86°	15	8	1 52 30	32
14	33.26°	50	27	1 51 06 40	32 24
15	31.89°	9	5	1 48	33 20

Figure 4. The proposed restorations at the beginning of the tablet according to the trigonometric, generating function, and reciprocal pair theories.

III contain the short sides and diagonals of right-angled triangles, then the values in the first column are \tan^2 or $1/\cos^2$ —and the table is arranged so that the acute angles of the triangles decrease by approximately 1° from line to line.

2. Neugebauer [19], and Aaboe following him, argued that the table was generated like this:

If p and q take on all whole values subject only to the conditions

- (1) $p > q > 0$,
- (2) p and q have no common divisor (save 1),

(3) p and q are not both odd, then the expressions

$$\begin{aligned} x &= p^2 - q^2 \quad [\text{our } s], \\ y &= 2pq \quad [\text{our } l], \\ z &= p^2 + q^2 \quad [\text{our } d], \end{aligned}$$

will produce all reduced Pythagorean number triples, and each triple only once [1, pp. 30–31].

The quest has then been to find how p and q were chosen.

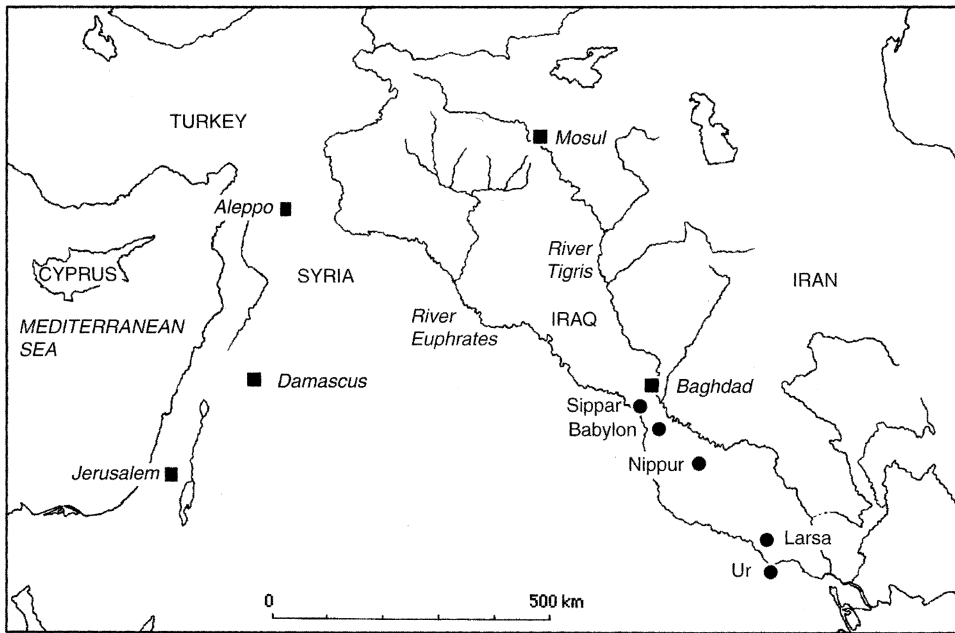


Figure 5. Map of the archaeological sites mentioned in the text. Drawing by the author.

3. Finally, the interpretation first put forward by Bruins [5], [6] and repeated in a cluster of independent publications about twenty years ago [4], [9], [30] is that the entries in the table are derived from reciprocal pairs x and $1/x$, running in descending numerical order from $2;24 \sim 0;25$ to $1;48 \sim 0;33\ 20$ (where \sim marks sexagesimal reciprocity). From these pairs the following “reduced triples” can be derived:

$$\begin{aligned} s' &= s/l = (x - 1/x)/2, \\ l' &= l/l = 1, \\ d' &= d/l = (x + 1/x)/2. \end{aligned}$$

The values given on the tablet, according to this theory, are all scaled up or down by common factors 2, 3, and 5 until the coprime values s and d are reached.

How are we to choose between the three theories? Internal mathematical evidence alone clearly isn't enough. We need to develop some criteria for assessing their historical merit. In general, we can say that the successful theory should not only be mathematically valid but historically, archaeologically, and linguistically sensitive too.

A great deal of emphasis has been laid on the uniqueness of Plimpton 322; how nothing remotely like it has been found in the corpus of Mesopotamian

mathematics. Indeed, this has been an implicit argument for treating Plimpton 322 in historical isolation. Admittedly we know of no other ancient *table* of Pythagorean triples, but Pythagorean triangles were a common subject for school mathematics problems in ancient Mesopotamia. This point has been made before (e.g., by Friberg [9]) but hasn't yet proved particularly helpful in deciding between the three interpretations of Plimpton 322. What we shall do instead is to make some *new* comparisons. None of the comparative material itself is new though: all but one of the documents I have chosen were published at the same time as Plimpton 322 or decades earlier.

First, Plimpton 322 is a table. Hundreds of other tables, both mathematical and nonmathematical, have been excavated from Mesopotamian archaeological sites. What can we learn from them?

Second, if Plimpton 322 is a *trigonometry* table, then there should be other evidence of measured angle from Mesopotamia. We shall go in search of this.

Third, Plimpton 322 contains words as well as numbers: the headings at the top of each column should tell us what the tablet is about. Some of the more difficult words also appear on other mathematical documents from Mesopotamia. Can they help us to understand their function on Plimpton 322?

Finally, Plimpton 322 was written by an *individual*. What, if anything, can we say about him or her, and why the tablet was made?

2 Turning the tables on generating functions

Let's start with some very general contextualisation. We can learn a lot about any tablet simply from its size, shape, and handwriting.

Plimpton 322 is named after its first Western owner, the New York publisher George A. Plimpton (see Donoghue [8]). He bequeathed his whole collection of historical mathematical books and artefacts to Columbia University in the mid-1930s along with a large number of personal effects. Surviving correspondence shows that he bought the tablet for \$10 from a well-known dealer called Edgar J. Banks in about 1922 [2]. Banks told him it came from an archaeological site called Senkereh in southern Iraq, whose ancient name was Larsa [Figure 5].

Vast numbers of cuneiform tablets were being illicitly excavated from Larsa at that time. Several big museums, such as the Louvre in Paris, Oxford's Ashmolean Museum, and the Yale Babylonian Collection bought thousands of them. Although Plimpton 322 doesn't look much like other *mathematical* tablets from Larsa, its format is strikingly similar to administrative tables from the area, first attested from the late 1820s BCE. The tablet YBC 4721 [12, no. 103], for example, is an account of grain destined for various cities within the kingdom of Larsa [Figure 6]. It was written in the city of Ur, then under Larsa's political control, in 1822 BCE and is now housed at Yale.

Like Plimpton 322 it is written on a "landscape" format tablet (that is, the writing runs along the longer axis) with a heading at the top of each column. Entries in the first column are sorted into descending numerical order. Calculations run from left to right across the table, while the final column lists the names of the officials responsible for each transaction. Although the scribes of Larsa mostly used the cuneiform script to write a Semitic language called Akkadian, they often used monosyllabic words from a much older language, Sumerian, as a kind of shorthand. Like the final column of Plimpton 322, the last heading on YBC 4721 carries the Sumerian writing MU.BI.IM for Akkadian *sumsu* ("its name"). Unlike Plimpton 322 though, the text is dated in the final line. There are about half a dozen published tables from the Larsa area with these same characteristics: all of them are dated to the short period 1822–1784 BCE and so, therefore, is Plimpton 322.

So we can already say that Plimpton 322 was written by someone familiar with the temple admin-

istration in the city of Larsa around 1800 BCE, at least twenty years before its conquest by Babylon in 1762. Sherlock Holmes, if he ever made it to Babylon, would have been over 100 miles away from the action: no ancient mathematics has ever been found there.

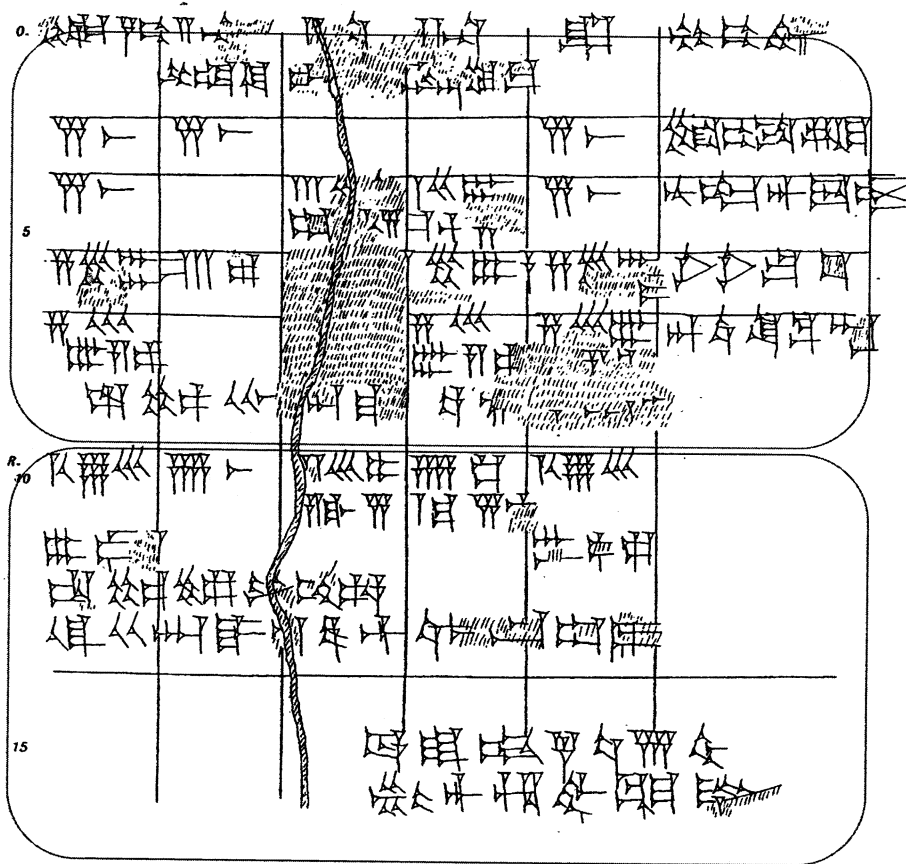
And the fact that Plimpton 322 follows the same formatting rules as all other tables from ancient Larsa leads us to dismiss Neugebauer's theory of generating functions. If the missing columns at the left of the tablet had listed p and q , they would not have been in descending numerical order and would thus have violated those formatting rules. Nor, under this theory, has anyone satisfactorily explained the presence of Column I in the table. There are other good reasons to eliminate the generating function theory; I deal with them in [29]. The trigonometry table and the reciprocal pairs remain.

3 Circling round trigonometry

We saw at the beginning of this article how differently from us the people of ancient Mesopotamia thought about triangles; that contrast with modern concepts runs right through their plane geometry.

For instance, YBC 7302 [20, p. 44] is roughly contemporary with Plimpton 322 [Figure 7]. From its circular shape and size (about 8 cm across) we know the tablet was used by a trainee scribe for school rough work. It shows a picture of a circle with three numbers inscribed in and around it in cuneiform writing: 3 on the top of the circle, 9 to the right of it, and 45 in the centre. Now 9 is clearly the square of 3, but what is the relationship of these numbers to 45? The answer lies in the spatial arrangement of the diagram. Looking closely, we can see that the 3 lies directly on the circumference of the circle, while the 45 is contained within it. The 9, on the other hand, has no physical connection to the rest of the picture. If on this basis we guess that 3 represents the length of the circumference and 45 the area of the circle, we should be looking for the relationship $A = c^2/4\pi$. We have the c^2 —that's the 9—and if we use the usual Mesopotamian school approximation $\pi \approx 3$, we get $A = 9/12$. This translates in base 60 to 45/60, namely, the 45 written in the circle.

When we teach geometry in school we have our students use the relationship $A = \pi r^2$; none of us, I would guess, would use $A = c^2/4\pi$ as a formula for the area of a circle. In modern mathematics the



<i>Grain debit</i>	<i>For Ur</i>	<i>For Mar-...</i>	<i>For</i>	<i>Total</i>	<i>Its name</i>
301 <gur>	301 <gur>			301 <gur>	Lipit-Suen
301 <gur>		214 gur, 285 sila	86 gur 15 sila	301 <gur>	Nur-Dagan
296 <gur>	180 gur	[60 gur]	56 <gur>	296 gur	Ili-eriba
277 gur 200 sila		[]	277 <gur> 200 sila	277 <gur> 200 sila	Samas-kima-ilisu
From 23 gur 40 sila of [Samas-...-aplu's] troops/workers.					
1176 gur 20 sila	481 <gur>	274 <gur> 485 sila	420 gur 95 sila	1176 gur 20 sila	
From the grain of Lu-am-...					
And from 23 gur 40 sila of Samas-...-aplu's troops/workers.					
Month 1, day 7,					
Year that Rim-Sin became king (1822 BCE).					

Figure 6. YBC 4721, after Grice [12, pl. XL]. 1 gur = 300 sila ≈ 300 litres.

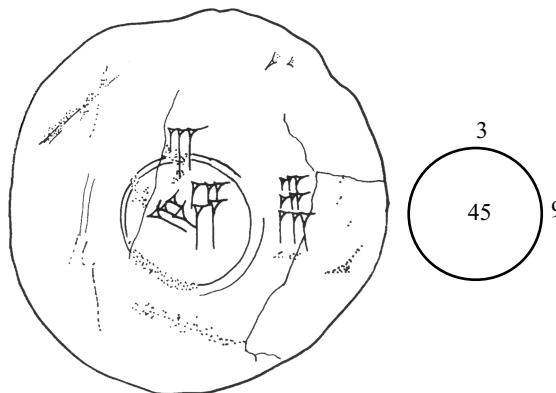


Figure 7. YBC 7302 (obverse). Drawing by the author.

circle is conceptualised as the area generated by a rotating line, the radius. In ancient Mesopotamia, by contrast, a circle was the shape contained within an equidistant circumference: note that there is no radius drawn on YBC 7302. There are many more examples of circle calculations from the early second millennium, and none of them involves a radius. Even when the diameter of a circle was known, its area was calculated by means of the circumference. We also see this conceptualisation in the language used: the word *kippatum*, literally “thing that curves,” means both the two-dimensional disc and the one-dimensional circumference that defines it. The conceptual and linguistic identification of a plane figure and one of its external lines is a key feature of Mesopotamian mathematics. For instance, the word *mithartum* (“thing that is equal and opposite to itself”) means both “square” and “side of square.” We run into big interpretational problems if we ignore these crucial terminological differences between ancient Mesopotamian and our own mathematics.

What does this tell us about Plimpton 322? That if plane figures were conceptualised, named, and defined from the inside out, then the centre of the circle and the idea of the rotating radius could not have played an important part in Mesopotamian mathematics. And if the rotating radius did not feature in the mathematical idea of the circle, then there was no conceptual framework for measured angle or trigonometry. In short, Plimpton 322 cannot have been a trigonometric table.

This should have been our intuition on later historical grounds anyway. Nearly two millennia after Plimpton 322 was written, Ptolemy conceptualised the circle as a diameter rotating about its centre in order to simplify his calculations of chords of arc—but those chords were functions of arc, not of angle (Toomer [31, p. 47]). (Ptolemy, working in Roman Egypt in the second century CE, was heavily reliant on Mesopotamian traditions: he used astronomical data from first millennium Assyria and Babylonia, adapted Mesopotamian mathematical methods, and even calculated in base 60.) Over the following millennium several generations of Indian and Iraqi scholars compiled tables of half-chords, but the conceptual transition from arc to angle was slow and halting.

Returning to the early second millennium BCE, I should emphasise two points. First, I do not mean that the ancient Mesopotamians *did not know* that circles could be generated by rotating radii. There is

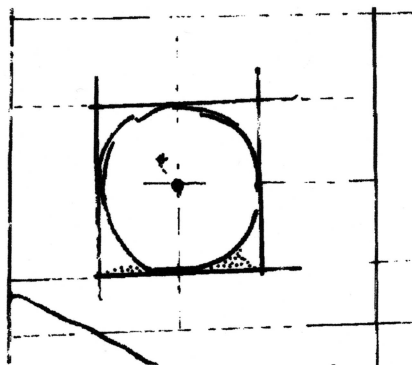


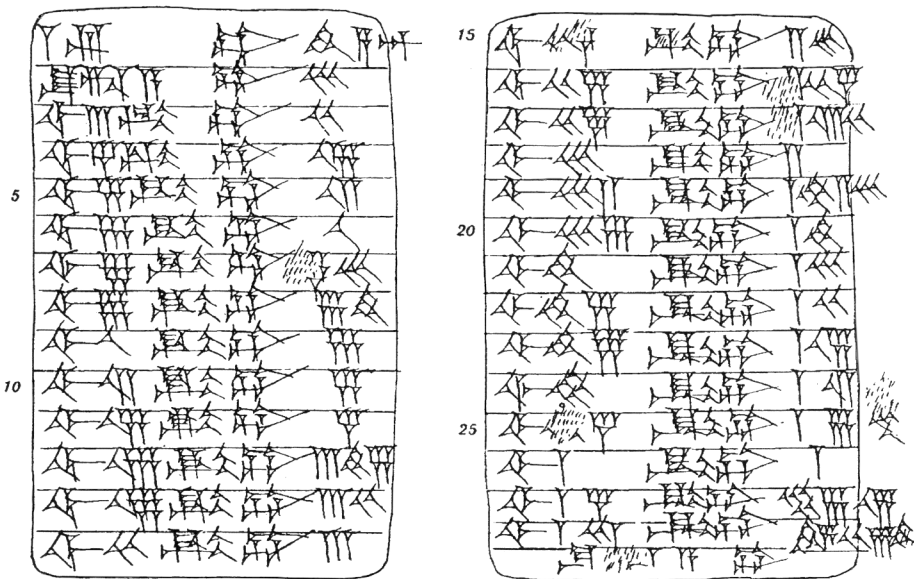
Figure 8. BM 15285 (detail). Drawing by the author [25, p. 214].

a great deal of visual evidence to show that they did. For example, BM 15285, a compilation of plane geometry problems from Larsa, depicts several circles whose deeply impressed centres reveal that they were drawn by means of rotating compasses [Figure 8]. But Mesopotamian mathematical concepts were as socially bounded as ours are: although we often draw circles free-hand without radii, even in mathematics classes, it would rarely cross our minds to teach our students $A = c^2/4\pi$. Equally, radii were known and used in ancient Mesopotamia, but played little part in the dominant outside-in conceptualisation of plane geometry. Second, neither do I mean that there was no concept of angle *at all* in ancient Mesopotamia. Gradients were used to measure the external slope of walls and ramps in formulations like “for every 1 cubit depth the slope (of the canal) is 1/2 cubit” (YBC 4666, rev. 25; [20, text K]). There was also a rough distinction made between right angles and what we might call “wrong angles,” namely, those configurations for which the Pythagorean rule held true or not, with probably a 10° to 15° leeway (see Robson [24]).

To sum up so far: the theory of generating functions is organisationally implausible, while the trigonometric theory is conceptually anachronistic. We are left with the theory of reciprocal pairs—how does it measure up to our historical expectations?

4 Words count too: reciprocal pairs

We can start by recognising that reciprocal pairs—unlike generating functions or trigonometry—played a key role in ancient Mesopotamian mathematics. Our best evidence is from the scribal schools of



Two thirds of 1 is 0;40.
 Its half is 0;30.
 The reciprocal of 2 is 0;30.
 The reciprocal of 3 is 0;20.
 The reciprocal of 4 is 0;15.
 The reciprocal of 5 is 0;12.
 The reciprocal of 6 is 0;10.
 The reciprocal of 8 is 0;07 30.
 The reciprocal of 9 is 0;06 40.
 The reciprocal of 10 is 0;06.
 The reciprocal of 12 is 0;05.
 The reciprocal of 15 is 0;04.
 The reciprocal of 16 is 0;03 45.
 The reciprocal of 18 is 0;03 20.
 The reciprocal of 20 is 0;03.

The reciprocal of 24 is 0;02 30.
 The reciprocal of 25 is 0;02 24.
 The reciprocal of 27 is 0;02 13 20.
 The reciprocal of 30 is 0;02.
 The reciprocal of 32 is 0;01 52 30.
 The reciprocal of 36 is 0;01 40.
 The reciprocal of 40 is 0;01 30.
 The reciprocal of 45 is 0;01 20.
 The reciprocal of 48 is 0;01 15.
 The reciprocal of 50 is 0;01 12.
 The reciprocal of 54 is 0;01 06 40.
 The reciprocal of 1 00 is 0;01.
 The reciprocal of 1 04 is 0;00 56 15.
 The reciprocal of 1 21 is 0;00 44 26 40.
 <Its half>

Figure 9. MLC 1670, after Clay [7, 37].

nineteenth and eighteenth century Larsa, Ur, and Nippur, where thousands of surviving practice copies show that scribal students had to learn their sexagesimal multiplication tables in the correct order and by heart. The first part of the series was the set of thirty standard reciprocal pairs encompassing all the sexagesimally regular integers from 2 to 81 (thereby including the squares of the integers 1 to 9) [Figure 9]. The trainees also learned how to calculate the reciprocals of regular numbers that were not in the standard list and practised division by means of finding reciprocals, as this was how all Mesopotamian divisions were carried out (see Robson [26, pp. 19–23]).

Looking at the reciprocals proposed as the starting point for Plimpton 322 [Figure 4], it turns out

that although only five pairs occur in the standard list, the other ten are widely in evidence elsewhere in Mesopotamian mathematics (as constants, for instance), or could be calculated trivially using methods known to have been taught in scribal schools. None of them is more than four sexagesimal places long, and they are listed in decreasing numerical order, thereby fulfilling our tabular expectations. But we haven't yet explained the purpose of the first surviving column: what are all those s^2/l^2 (or d^2/l^2) doing there?

The headings at the top of the table ought to tell us: that, after all, is their function. We have already seen that the last column, which contains only a line-count, is headed like the other tables from Larsa with the signs MU.BI.IM meaning *šumšu* ("its name").

The two columns immediately preceding that, we remember, contain what we can conveniently think of as the shortest sides and diagonals of right-angled triangles. They are headed IB.SI_8 SAG and IB.SI_8 *ṣi-li-ip-tim* for *mitharti pūtim* and *mitharti ṣilipitim* (“square of the short side” and “square of the diagonal,” respectively). This contradiction disappears when we recall that Mesopotamian plane figures are defined and named for their key external lines. We can thus adjust our translations to read “square-side” of the short side and diagonal, respectively. (I am using the translation “diagonal” here rather than “hypotenuse” to indicate that this is a general word for the transversal of a figure, not restricted to triangles.)

Last and by far the most difficult, the heading of the first surviving column reads

[*ta*]-*ki-il-ti ṣi-li-ip-tim*

[*sa* 1 *in*]-*na-as-sà-ḥu-ú-ma* SAG *i-il-lu-ú*

(for Akkadian *takilti ṣilipitim sa istēn innassaḥūma pūtum illū*),

where square brackets mark missing cuneiform signs that I have restored. Surprisingly, no one has been able to improve convincingly on the translation made by Neugebauer and Sachs when they first published Plimpton 322 [20, p. 40]. They were uncertain about the first word and the last word, as well as what was missing at the beginning of the second line. In fact the last word is legible, if a little squashed. The breaks at the beginnings of the lines can be filled in, and the whole understood, through comparison with other mathematical documents that use the same terminology. We end up with something like this:

The *takiltum*-square of the diagonal from which 1 is torn out, so that the short side comes up.

To understand what exactly that means, and how it relates to the reciprocal pairs, we need to look at one more mathematical tablet, YBC 6967 [20, text Ua]. This tablet is almost certainly from late nineteenth to early eighteenth century Larsa, like Plimpton 322. It contains instructions for solving a school problem about reciprocal pairs. As Jens Høyrup has shown, we can best understand this sort of mathematics not as algebra but as a very concrete cut-and-paste geometry [13, pp. 262–266]. Once again square brackets show restorations of missing text.

[A reciprocal] exceeds its reciprocal by 7. What are [the reciprocal] and its reciprocal?

The product of the mystery reciprocals is by definition 1 (or any power of 60). The fact that their difference is an integer suggests that we should think

of them as integers too. We can thus conceptualise them as the unknown lengths of a rectangle with area 60 [Figure 10].

You: break in half the 7 by which the reciprocal exceeds its reciprocal, and 3;30 (will come up). Multiply 3;30 by 3;30 and 12;15 (will come up).

Following the instructions, we can move the broken piece of the rectangle to form an L-shaped figure, still of area 60, around an imaginary square of area $12 \frac{1}{4}$.

Append [1 00, the area,] to the 12;15 which came up for you and 1 12;15 (will come up). What is [the square-side of 1] 12;15? 8;30.

Together, therefore, they comprise a large square of area $72 \frac{1}{4}$ and side $8 \frac{1}{2}$.

Put down [8;30 and] 8;30, its equivalent, and subtract 3;30, the *takiltum*-square, from one (of them); append (3;30) to one (of them). One is 12, the other is 5. The reciprocal is 12, its reciprocal 5.

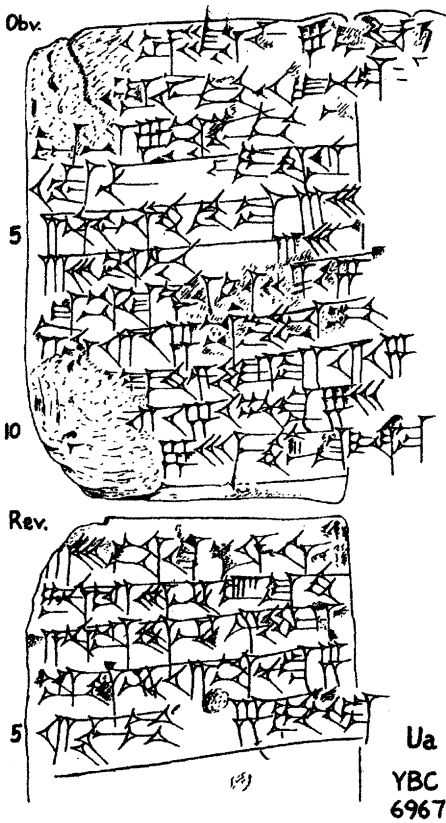
We remove the vertical side of the imaginary small square from that of the large composite square, reverting to the smaller side of the original rectangle, a side whose length is 5. We find the longer side of the rectangle by adding the horizontal side of the imaginary square onto that of the large composite square and arrive at the answer 12.

If instead we choose a reciprocal pair whose product is not 60 but 1, their product can be imagined as a much longer, narrower rectangle than in Figure 10. But the semidifference of the reciprocals, $(x - 1/x)/2$, can still be found and the rectangle rearranged to form an L-shaped gnomon, still of area 1. Its outer edges will still be the lengths of a large square, and its inner edges the lengths of a small square. That is, we will have a composite large square that is the sum of 1 (itself a square) and an imaginary small square. This set of three squares, all generated by a pair of reciprocals, obeys the Pythagorean rule $d^2 = s^2 + l^2$. Their sides, in other words, are the Pythagorean triple we have been looking for.

Let us look again at the heading of Column I:

The *takiltum*-square of the diagonal from which 1 is torn out, so that the short side comes up.

It describes the area of the large square, composed of 1 plus the small square—the verb *ilam* (“come up”), we have seen, is the standard term for “to result.” Our restoration dilemma is now solved: we should



[A reciprocal] exceeds its reciprocal by 7.
 What are [the reciprocal] and its reciprocal?
 You: break in half the 7 by which the reciprocal exceeds its reciprocal, and 3;30 (will come up).
 Multiply 3;30 by 3;30 and 12;15 (will come up).
 Append [1 00, the area,] to the 12;15 which came up for you and 1 12;15 (will come up).
 What is [the square-side of 1] 12;15? 8;30.
 Put down [8;30 and] 8;30, its equivalent, and subtract 3;30, the *takiltum*-square, from one (of them); append (3;30) to one (of them).
 One is 12, the other is 5.
 The reciprocal is 12, its reciprocal 5.

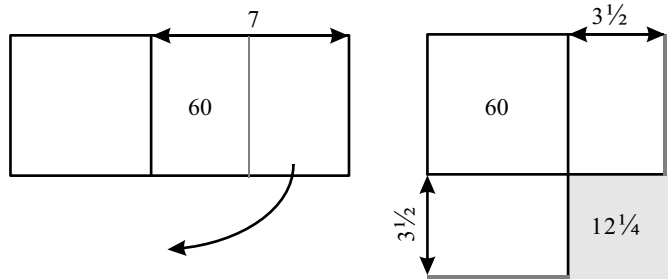


Figure 10. YBC 6967, after Neugebauer and Sachs [20, pl. 17].

put 1s at the beginning of every entry. There is one small terminological discrepancy left to deal with: in Plimpton 322 *takiltum* refers to the area of the large composite square, while in YBC 6967 it means the side of the small imaginary square. We know by now to expect squares and their sides to be named identically so that is not a problem. The word itself, a technical derivative of the verb *kullum* (“to multiply lengths together into areas”) does not suggest that its meaning should be restricted to either the little square or the big square but that its pattern of attestation is restricted to exactly these cut-and-paste geometrical scenarios.

We have found, then, the most historically, culturally, and linguistically convincing of our three interpretations of Plimpton 322: a list of regular reciprocal pairs, each four places long or shorter, was drawn up in the usual decreasing numerical order on the missing part of the tablet. They were used to find the short sides *s* and diagonals *d* of triangles with long sides of length $l = 1$ by the method of completing the square. One of the intermediate results was recorded in the first extant column. Then common

factors were eliminated from the triples produced to give the coprime short sides and diagonals listed in Columns II and III.

All we need to know now is who wrote Plimpton 322 and for what purpose—but that is easier said than done!

5 In search of an author

Ancient Mesopotamia was a culture that prized anonymised tradition over individual creativity. Even the greatest works of literature were attributed to deities or to long-dead historical figures (see Michalowski [17]). It is very unlikely that we will ever be able to put a name to our author, let alone outline his or her personality or life history. We can find out a great deal of more general information though. For instance, it is virtually certain that our author was male: all the known female scribes from ancient Mesopotamia lived and worked much further north, in central and northern Iraq. We can also rule out the possibility that our author was a mathematician in either of the senses we normally mean. He

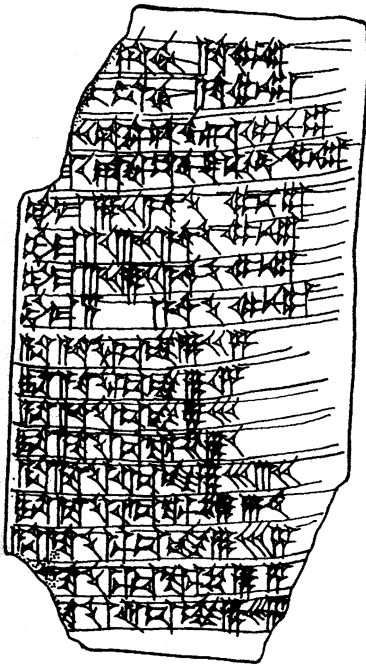


Figure 11. BM 80209 (obverse). Drawing by the author.

[If each] square side is [...], what is the area?
 [If each] square side is [...], what is the area?
 [If] each square side is 20, what is the diagonal?
 [If] each square side is 10, what is the border?
 If the area is 8 20, what is the circumference?
 If the area is 2 13 20, what is the circumference?
 If the area is 3 28 20, what is the circumference?
 If the area is 5, what is the circumference?
 To the area of the circle add $\frac{1}{2}$ a length: 8 25.
 From the area of the circle take $\frac{1}{2}$ a length: 8 15.
 To the area of the circle add 1 length: 8 30.
 From the area of the circle take 1 length: 8 10.
 To the area of the circle add $1\frac{1}{3}$ lengths: 8 33 20.
 From the area of the circle take $1\frac{1}{3}$ lengths: 8 06 40.
 To the area of the circle add $1\frac{1}{2}$ lengths: 8 35.
 [From] the area of the circle take $1\frac{1}{2}$ lengths: 8 05.
 [To] the area of the circle add $1\frac{2}{3}$ lengths: 8 36 [40].

cannot have been a *professional* mathematician—the professionalisation of academic disciplines is a phenomenon of the very recent past. Nor was he likely to have been an *amateur* mathematician like those of Classical Antiquity and the Middle Ages, i.e., an educated member of the merchant classes or ruling elite for whom wealth, high status, or royal patronage provided enough leisure time to indulge his mathematical inclinations [18, ch. 7]. There is not one example of this type of individual in the whole of Mesopotamia’s three-thousand year history. Rather, he must have been someone who used literacy, arithmetic, and mathematical skills in the course of his working life.

We can say something more positive about the author’s identity by recalling some of our earlier conclusions. First, the methods used to construct Plimpton 322—reciprocal pairs, cut-and-paste geometry, completing the square, dividing by regular common factors—were all simple techniques taught in scribal schools. Our author could have been a trainee scribe or a teacher. Second, he was familiar with the format of documents used by the temple and palace administrators of Larsa. That rules out the option that he was a student, but indicates instead that he was a professional bureaucratic scribe. In that case he would have been highly numerate, for the vast majority of ancient administrative documents related

to quantity surveying or accountancy. If the author of Plimpton 322 was a teacher, then he was almost certainly a bureaucrat too: we know the names and primary professions of about half a dozen ancient Mesopotamian teachers, and all of them had careers in temple administration.

It is highly unlikely, however, that Plimpton 322 was written for the temple bureaucracy: its organisational structure most closely resembles a class of school mathematics documents that we might call “teachers’ problem lists.” A good example is BM 80209, originally from ancient Sippar near modern Baghdad but now housed in the British Museum [10]. It repeats a few school mathematics problems over and over, each time giving a different set of numerical data that will yield a tidy integer answer [Figure 11]. Plimpton 322 is also a repetition of the same mathematical set-up fifteen times, each with a different group of well-behaved regular numbers. It would have enabled a teacher to set his students repeated exercises on the same mathematical problem, and to check their intermediate and final answers without repeating the calculations himself.

6 Conclusions

I stated at the beginning that this paper would be both about Plimpton 322 and about historical

methods more generally. A great deal of the history of mathematics concerns periods, languages, and settings that we know a lot about and share common ground with: we are already more or less familiar with Galois's cultural background, for instance, or Newton's. We are also helped enormously by knowing their identities, their life histories, other writings by them and their contemporaries. This allows us to contextualise the mathematical content of their work, helping us to understand it as they did. But when we start to study mathematics from cultures whose languages, social practices, and common knowledge we do not share, we have to work considerably harder at positioning it within a historically and mathematically plausible framework.²

Plimpton 322, analysed solely as a piece of mathematics, looked very modern, although it was impossible to say which branch of modern mathematics it most closely resembled: trigonometry, number theory, or algebra. It seemed millennia ahead of its time, incomparably more sophisticated than other ancient mathematical documents. But if we treat Plimpton 322 as a cuneiform tablet that just happens to have mathematics on it, a very different picture emerges. We see that it is a product of a very particular place and time, heavily dependent on the ancient scribal environment for its physical layout as a table, its mathematical content, and its function as a teacher's aid. All the techniques it uses are widely attested elsewhere in the corpus of ancient Mesopotamian school mathematics. In this light we can admire the organisational and arithmetical skills of its ancient author but can no longer treat him as a far-sighted genius. Any resemblance Plimpton 322 might bear to modern mathematics is in our minds, not his.

References

1. A. Aaboe, *Episodes from the Early History of Mathematics*, New Mathematical Library, vol. 13, Mathematical Association of America, Washington, D.C., 1964.
2. E. J. Banks, Description of four tablets sent to Plimpton, unpublished manuscript, Columbia University Libraries Special Collections, Cuneiform Collection, no date.
3. D. R. Brown, *Mesopotamian Planetary Astronomy-Astrology*, Cuneiform Monographs, vol. 18, Styx Publications, Groningen, 2000.
4. R. C. Buck, Sherlock Holmes in Babylon, *Amer. Math. Monthly* 87 (1980), 335–345.
5. E. M. Bruins, On Plimpton 322, Pythagorean numbers in Babylonian mathematics, *Koninklijke Nederlandse Akademie van Wetenschappen Proceedings* 52 (1949), 629–632.
6. —, Pythagorean triads in Babylonian mathematics: The errors on Plimpton 322, *Sumer* 11 (1955), 117–121.
7. A. T. Clay, *Babylonian Records in the Library of J. Pierpont Morgan*, vol. 4, Yale University Press, New Haven, 1923.
8. E. F. Donoghue, In search of mathematical treasures: David Eugene Smith and George Arthur Plimpton, *Historia Mathematica* 25 (1998), 359–365.
9. J. Friberg, Methods and traditions of Babylonian mathematics: Plimpton 322, Pythagorean triples and the Babylonian triangle parameter equations, *Historia Mathematica* 8 (1981), 277–318.
10. —, Methods and traditions of Babylonian mathematics, II: an Old Babylonian catalogue text with equations for squares and circles, *Journal of Cuneiform Studies* 33 (1981), 57–64.
11. —, Mathematik, In *Real Lexikon der Assyriologie*, vol. 7, D. O. Edzard, ed., De Gruyter, Berlin/Leipzig, 1987–90, pp. 531–585 (in English).
12. E. M. Grice, *Records from Ur and Larsa dated in the Larsa Dynasty*, Yale Oriental Series, Babylonian Texts, vol. 5, Yale University Press, New Haven, 1919.
13. J. Høyrup, Algebra and naive geometry. An investigation of some basic aspects of Old Babylonian mathematical thought, *Altorientalische Forschungen* 17 (1990) 27–69; 262–354.
14. —, *In Measure, Number, and Weight: Studies in Mathematics and Culture*, State University of New York Press, Albany, 1994.
15. D. E. Joyce, Plimpton 322, <<http://aleph0.clarku.edu/~djjoyce/mathhist/plimpnote.html>>, Clark University, 1995.
16. A. Kuhrt, *The ancient Near East: c. 3000–300 BC*, Routledge History of the Ancient World, 2 vols., Routledge, London, 1995.
17. P. Michalowski, Sailing to Babylon, reading the dark side of the moon, in *The Study of the Ancient Near East in the Twenty-First Century: the William Foxwell Albright Centennial Conference*, J. S. Cooper & G. M. Schwartz, eds., Eisenbrauns, Winona Lake, IN, 1995, pp. 177–193.

²For the social history of Mesopotamian mathematics, see Nissen et al. [21], Høyrup [14], and Robson [25, pp. 138–183], [27], [28], all with further bibliography. Friberg [11] is a comprehensive survey of Mesopotamian mathematical techniques. For the social history of Mesopotamia in general, see Walker [33], Roaf [23], Postgate [22], Kuhrt [16], and Van De Mieroop [32].

18. R. Netz, *The Shaping of Deduction in Greek Mathematics: a Study in Cognitive History*, Ideas in Context, vol. 51, Cambridge University Press, Cambridge, 1999.
19. O. Neugebauer, *The Exact Sciences in Antiquity*, Munksgaard, Copenhagen, 1951; 2nd ed., Brown University Press, Princeton, NJ, 1957; reprint ed., Dover, New York, 1969. All page references are given according to the Dover edition.
20. — and A. J. Sachs, *Mathematical Cuneiform Texts*, American Oriental Series, vol. 29, American Oriental Society and the American Schools of Oriental Research, New Haven, 1945.
21. H. Nissen, P. Damerow, and R. K. Englund, *Archaic Bookkeeping: Early Writing and Techniques of Economic Administration in the Ancient Near East*, University of Chicago Press, Chicago/London, 1993.
22. J. N. Postgate, *Early Mesopotamia: Society and Economy at the Dawn of History*, Routledge, London/New York, 1992.
23. M. Roaf, *Cultural Atlas of Mesopotamia and the Ancient Near East*, Facts on File, London, 1990.
24. E. Robson, Three Old Babylonian methods for dealing with “Pythagorean” triangles, *Journal of Cuneiform Studies* 49 (1997), 51–72.
25. —, *Mesopotamian Mathematics, 2100–1600 BC: Technical Constants in Bureaucracy and Education*, Oxford Editions of Cuneiform Texts, vol. 14, Clarendon Press, Oxford, 1999.
26. —, Mathematical cuneiform tablets in Philadelphia, part I: problems and calculations, *SCIAMVS—Sources and Commentaries in Exact Sciences* 1 (2000), 11–48.
27. —, Mesopotamian mathematics: some historical background, in *Using History to Teach Mathematics*, V. J. Katz, ed., Mathematical Association of America, Washington, D.C., 2000, pp. 149–158.
28. —, The uses of mathematics in ancient Iraq, 6000–600 BC, in *Mathematics Across Cultures: the History of Non-Western Mathematics*, H. Selin, ed., Kluwer Academic Publishers, Dordrecht, 2000, pp. 93–113.
29. —, Neither Sherlock Holmes nor Babylon: a re-assessment of Plimpton 322, *Historia Mathematica* 28 (2001), 167–206.
30. O. Schmidt, On Plimpton 322. Pythagorean numbers in Babylonian mathematics, *Centaurus* 24 (1980), 4–13.
31. G. J. Toomer, *Ptolemy’s Almagest*, Duckworth’s, London, 1984.
32. M. Van De Mieroop, *Cuneiform Texts and the Writing of History*, Routledge, London and New York, 1999.
33. C. B. F. Walker, *Cuneiform*, British Museum Press, London, 1987.

Mathematics, 600 B.C.–600 A.D.

MAX DEHN

American Mathematical Monthly 50 (1943), 357–360;
50 (1943), 411–414; 51 (1944), 25–31; 51 (1944), 149–157

1 600 B.C.–400 B.C.

1.1 Introduction

Isolated arithmetical and geometrical facts were, without doubt, known in prehistoric times much as such facts are now known among the most primitive tribes. Rather advanced mathematical knowledge appears in ancient Egyptian papyri (for instance in the Rhind Papyrus of the 14th century B.C.) and on numerous Babylonian cuneiform texts dating from 2000 B.C. onwards. Certainly the Greeks learned many of the algebraic methods and the techniques of geometric measurements from these ancient peoples through the lively commerce of the Eastern Mediterranean. Our reports begin with Greek mathematics after 600 B.C.

1.2 Sources

The sources for the history of mathematics in Greece during the period from 600 B.C. to 400 B.C. are very scarce and unreliable. We have a fragment of mathematical history by *Eudemos* (ca. 320 B.C.) in an excerpt of the sixth century A.D. This fragment itself is in a bad state, corrupted by later changes. There are, however, scattered among the works of Greek authors, enough passages concerned with the mathematics and mathematicians of Ancient Greece, for us to derive a fairly clear idea of this early period.

1.3 Early Greeks

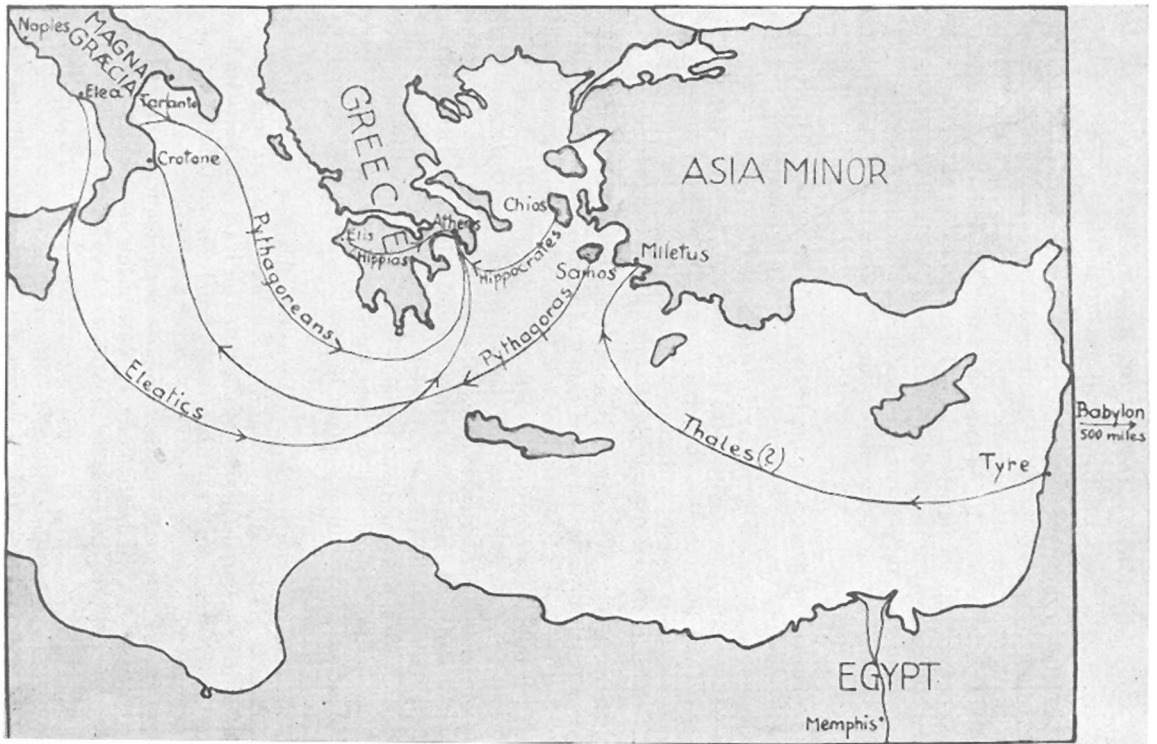
While there is no mathematician known from ancient Egypt or Babylon, we do know the names of famous Greek mathematicians.

Thales (ca. 600 B.C.) of Miletus (see map), who was probably of Phoenician origin, is known as the father of Greek mathematics. He had many disciples. It may be that there is a direct connection between him and *Pythagoras* (ca. 550 B.C.) from Samos. The latter, who was the head of an aristocratic brotherhood, a school of wisdom and science, was a political and philosophical leader in Southern Italy. He emphasized the importance of mathematics in the higher or liberal education, and for many centuries his name invoked an aura of mysticism. After his death, his school flourished for more than a hundred years, and numbered several famous mathematicians among its members. They will be mentioned in the second section.

Hippocrates of Chios (ca. 450 B.C.) was probably not connected with the Pythagoreans. He taught mathematics at Athens. We have a fragment of his mathematical work transmitted by *Eudemos*. This is the first *published* mathematical investigation known. *Hippocrates* is probably also the author of the first manual of geometry.

Hippias of Elis (ca. 430 B.C.) was a famous Sophist, a man with vast knowledge in mathematics and astronomy. An outstanding teacher, he was paid for his courses, which he gave mainly at Athens, where teaching and research in mathematics were concentrated at the end of the period with which the present report is concerned.

Even in this early period we begin to see many of the features of modern scientific activity: authors famous for their achievements, ambitious to find new results; renowned teachers; pupils eager to learn; books where results are collected, digested, and presented in such a way that the reader understands



the facts and proofs, and is inspired to do research himself.

1.4 Achievements

What is left to us of Babylonian and Egyptian mathematics shows only prescriptions for computations or for solutions of particular problems. But we find in the old Greek mathematics, proofs of the given solutions of problems and of the various theorems; we find convincing explanations. Great problems are proposed and treated. Problems of construction are solved with the help of ruler and compass. Among such problems are the conversion of areas into each other (see Figure 1), the most important case being the squaring of the rectangle; and the construction of the regular pentagon by means of the golden ratio.

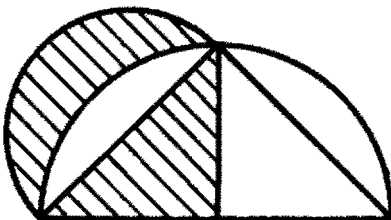


Figure 1.

Also propounded at this period were three classical problems of construction: the squaring of the circle, the trisection of an angle, and the duplication of the cube. The first two problems stimulated the *construction of the first curve* apart from the “naturally” given circle. This curve, which was invented by Hippias, is the quadratrix, whose equation in rectangular coordinates is $y = x \cot(\pi x/2r)$. The construction of points on this curve, approaching the y -axis at the level $y = 2r/\pi$, corresponds to the Archimedean computation of the perimeters of regular polygons approaching the circumference of a circle. In the construction of Hippias, the limiting process is visualized by the continuous curve approaching the y -axis.

The problems of trisecting an angle and of duplicating the cube led to new mechanical devices other than the compass, and finally, at the beginning of the next period, resulted in the discovery of the conics.

1.5 Theorems

Probably the first theorems, found by the Greeks, were propositions about angles. The *Pythagorean theorem*, as a relation between the lengths of the sides of a right triangle, was in all likelihood already known to the Babylonians; as a theorem about

areas it is perhaps a Greek achievement. At all events, the knowledge of this theorem which was always attributed to Pythagoras himself, was a matter of great moment to all educated Greeks.

Endeavoring to square the circle, Hippocrates discovered areas bounded by two circular arcs which could be constructed by compass and rule. The simplest case is indicated in Figure 1, in which the two shaded areas are equal.

To this period belongs the discovery and the construction of the *five regular solids*.

The greatest achievement of this epoch was the discovery and proof of the existence of *irrational ratios* in the incommensurability of side and diagonal of a square. Whether the original proof was given by arithmetical or geometrical methods is unknown. This was the first example of a mathematical truth contrary to naively simplifying intuition. The necessity for a strict proof became apparent, and this influenced the whole development of mathematics in the direction of rigor.

Further, we have the discovery of the projection of the infinite process of counting into arithmetical, geometrical, and kinematic ideas. These phenomena were found and discussed by the *Eleates* (Elea, a city of Southern Italy). The finite sum of an infinite geometric progression, the indefinite subdivision of a finite line or of a finite movement were all in contradiction to naive intuition and provoked profound problems as well as new constructions in Philosophy.

2 400 B.C.–300 B.C.

2.1 Survey of the century

The most important men of this period are Plato and Aristotle. They clarified the aims and methods of scientific work. Not only did they dominate the spiritual life of this era, but they have remained to the present day—at different times one more than the other—the leaders of all people struggling to find the truth and to order the world of phenomena.

We owe to this period the outstanding systematic work on mathematics by Euclid (ca. 300 B.C.). It was used as a textbook soon after it was written, superseding all other textbooks written before it. Euclid's work was the only textbook for the elements of mathematics everywhere until about one hundred and fifty years ago and is even used in some countries today (for example, England).

A little older than Euclid's *Elements* is the oldest mathematical treatise preserved in its original

form—a work of Autolycus belonging to the domain of applied mathematics which describes the simplest phenomena of the movement of the stars as phenomena in the geometry of the sphere.

The Academy founded by Plato about 380 B.C. at Athens favored the study of mathematics. Important progress was made at the Academy in both mathematical method and mathematical knowledge. Typical scholarly work was done in the Peripatetic School founded by Aristotle about 350 B.C. at Athens. Eudemus, a member of this school, was the author of the first history of mathematics. For the larger part of this century Athens was the center of mathematics. However, at the end of the period, in line with political developments, Alexandria, the city founded in Egypt in 331 B.C. by Alexander the Great, became an important cultural and especially mathematical center. Euclid taught here.

2.2 Mathematical reasoning

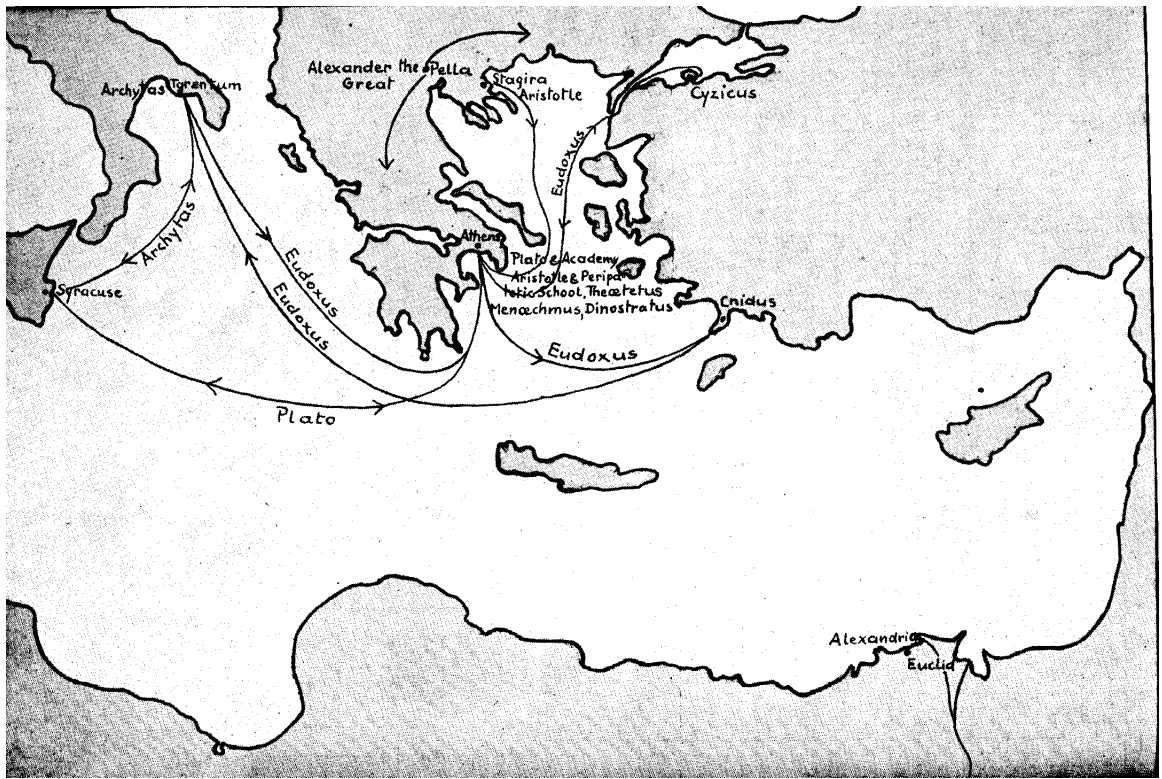
The philosophical discussions of this period led the mathematician to a higher level of consciousness of what he was doing. He became aware that the objects of his geometrical research had no existence in the outer reality appearing to our senses. They are something between this reality and the realm of ideas to which such concepts as that of the integers belong.

Further he was taught that it was his duty to formulate the foundations of his deductions, the definitions as well as the basic suppositions (the axioms). Also the form of the deduction itself was strongly influenced by the philosophical discussions.

The analytical method was introduced. This method starts with the assumption that the required construction has been carried out, and so leads to simpler figures easier to construct. Thereafter, one returns to the original problem of construction.

Connected with this method is the method of indirect proofs of theorems, which was probably already used in the first period, but now was in a certain sense the fashion in mathematical works. Aristotle laid the logical foundations upon which this method rests, the axiom of contradiction and the axiom of the excluded third. Both devices have turned out to be of great importance in modern discussions of mathematical methods.

A special case of the indirect proof appears in the method of exhaustion, indispensable for the proofs of theorems concerning areas not bounded by straight lines or volumes of general polyhedra (for example, pyramids).



2.3 Various achievements

A great achievement was the invention of a sound method of handling irrational ratios: this was accomplished by embedding them in the set of the rational ratios. This method is developed to a high degree of perfection in the Fifth Book of Euclid's *Elements*. The two mathematicians Theatetus of Athens and Eudoxus of Cnidus, both intimately connected with Plato, certainly contributed a great deal to this development.

To this time belongs, so far as we know, the discovery of the simplest properties of the conics. Menaechmus, also a follower of Plato, is believed to be the discoverer of all three types of conics, regarding them as the loci consisting of the intersection of a cone with a plane perpendicular to the generating line of the cone. Menaechmus used the hyperbola and the parabola simultaneously for the solution of the problem of doubling the cube, that is to construct $\sqrt[3]{2}$.

A little older than Menaechmus, Archytas of Tarentum used a three-dimensional construction for the duplication of the cube. He must have been a man of extraordinary scientific renown. The Roman poet Horace wrote a poem about him, but unfortunately

associated his name incorrectly with the achievements of Archimedes.

Menaechmus' brother, Dinostratus, in squaring the circle by Hippias' quadratrix, proved

$$\lim x \cot \frac{x\pi}{2r} = \frac{2r}{\pi}.$$

His proof is exact in the modern sense under the assumption that the quadratrix is a continuous curve.

2.4 Euclid's *Elements*

It is to the end of this period that we must assign the *Elements* of Euclid, probably written in Alexandria. We find there the greatest part of what nowadays is called elementary geometry. Some important elementary theorems, such as those concerning the intersection of the medians or the altitudes of a triangle, are not to be found there.

Euclid's final aim was obviously the metric theory of the regular solids. In this theory occur various irrational ratios. This gave Euclid the opportunity to build on a broad basis the theory of the domain of irrationals which contains as special cases the irrationals associated with the regular solids. Thus he first developed the theory of the whole numbers.

Here is found the process for determining the greatest common divisor of two whole numbers. This process, called the Euclidean algorithm, dominates under many different guises both the elementary and the advanced theories of arithmetic and algebra. He developed also other theories of purely arithmetical interest. We cite as an example the theory of perfect numbers, which are defined by the property that each is equal to the sum of its divisors (e.g. 6, 28, 496). This theory has not made much progress since the time of Euclid.

Then follows the comprehensive theory of those irrationals which are generated by using, apart from addition and multiplication, the single or double extraction of a square root. Such irrationals occur in the metric theory of the regular solids. An example of this occurrence is found in the fact that the side of a regular pentagon is equal to $\sqrt{10 - 2\sqrt{5}}r/2$, where r is the radius of the circumscribed circle. We find here no attempt to determine rational approximants to these irrationals; Euclid's main concern was to determine algebraic relations between the occurring irrationals.

Of great importance is the introduction of the postulate of parallels in the beginning of Euclid's work. It is known from remarks of Aristotle that the theory of parallels worried mathematicians. The introduction of a theorem about parallels as a postulate was an audacious device. It made possible the rigorous construction of this geometry, but caused much trouble to mathematicians through the ages until modern times.

Beside the *Elements*, Euclid wrote other works which are for the greatest part only preserved in fragments. Of importance for this review is his book on *Porisms*, some fragments of which we find with Pappus' work. (Pappus lived more than five hundred years after Euclid.) In this work, Euclid probably approached problems concerning functions, especially linear functions and their geometrical equivalents as embodied in straight lines, circles, and pencils of straight lines passing through a common point.

3 300 B.C.–200 B.C.

3.1 The life of Archimedes

The third century before Christ was dominated by the achievements of two men, Archimedes and Apollonius.

Archimedes has been considered, during his life and since his death up to this very day, the best

known mathematician. He died in the year 212 B.C. when the Roman army conquered Syracuse. About fifty years later, the Greek historian Polybius, living in Rome and writing on Roman history, related how Archimedes, by his mechanical inventions, became a formidable adversary to the Roman general who besieged Archimedes' native city. "The soul of one man," writes Polybius, "created almost insurmountable obstacles for the Roman army."

Archimedes saw the mathematical structure of static phenomena. This insight enabled him to do important engineering work in war and peace, and to bring to light the basic principles of the statics of rigid bodies and ideal fluids. He does not seem, however, to have attached too high a value to these achievements, since he wanted on his own tombstone a figure symbolic of his measurement of the sphere: a sphere with the circumscribed cylinder.

He was of noble birth, a kinsman and friend of Hiero, king of Syracuse. Syracuse had been for many centuries a center of commerce and culture, a point where Greeks and Phoenicians met. It is very probable that Archimedes went to Alexandria and studied there with the successors of Euclid.

A great part of his works is extant. There are no textbooks among them; many of them are concerned with theories or single problems. He was well aware of the value of discoveries and claimed the priority for his own. To show his superiority, he proposed to his Alexandrian competitors certain problems to be solved and certain theorems to be proved, intentionally including among the theorems some wrong ones. To take pride in one's own discoveries is perhaps not the sign of a philosophical mind, but it is certainly characteristic of times of great scientific progress. Archimedes was primarily a research man. He founded no school and had, as far as we know, no personal followers.

Archimedes died in the year 212 B.C. at the hand of a common Roman soldier. His tomb was found and restored by Cicero one hundred and fifty years later, when the latter was governor of Sicily. The Roman statesman did not share our reverent feeling for the unique genius of Archimedes, but spoke with condescending pity of the "modest man operating with sand and writing stylus" (paper and pen).

3.2 The achievements of Archimedes

Foundations of mathematics. He was probably the first to emphasize the axiomatic foundation of continuity, stating the following postulate of basic impor-



tance for all non-algebraic operations: the difference of two unequal quantities of the same kind, when added to itself a sufficient number of times, will exceed any other quantity of the same kind. This postulate was called, in the nineteenth century, the postulate of Archimedes. But Archimedes himself did not call it an axiom. He says only that he had to assume it for his deductions and that the mathematicians before him, by tacitly using it, had achieved results universally acknowledged as right. This attitude again shows Archimedes' non-philosophical trend of mind. He did not concern himself with eternal truths and ideas; he preferred to reach his aim by assuming the obvious as true.

Of similar character was the other assumption enabling him to assign to a plane (convex) curve a length, and to a curved (convex) surface an area. We may formulate the assumption this way: if one closed convex surface (curve) is completely inside another closed convex surface (curve), then the former has a smaller area (length) than the latter. It would be interesting to know how he would have assigned a length to a non-plane curve or an area to a surface of negative Gaussian curvature; for example, to the hyperboloid of one sheet.

An application of the "Archimedean" axiom is the computation of the number of grains of sand suffi-

cient to fill the "universe" in the sense of the astronomers of his time. He generated this number by means of exponential operations. This semi-popular work, addressed to Hiero's son Gelo, shows the surprising power of mathematical symbols.

An application of the assumption concerning the length of a plane curve was the approximation of π , the ratio of the circumference of a circle to the diameter. We have seen that, in Euclid, there are no approximations of irrational numbers.

Archimedes computed the length of the perimeter of the regular polygons of 6×2^n sides, and in this manner obtained (in modern notation)

$$2^n \frac{a_n}{b_n} > \frac{\pi}{6} > 2^{n-1} a_n, \quad (n = 1, 2, \dots),$$

where

$$b_n = \sqrt{2 + b_{n-1}},$$

$$a_n = \sqrt{2 - b_{n-1}},$$

with

$$b_0 = \sqrt{3}.$$

Archimedes gives approximate values for the resulting algebraic numbers for $n = 1, 2, 3, 4$ and obtains

$$\frac{22}{7} > \pi > \frac{223}{71}.$$

The history of the measurement of the circle goes back to very ancient times. The number 3 as an approximate value of π is used by people of low scientific standing. The comparison between the simple experiment yielding this approximation and Archimedes' procedure enabling us to determine π to any degree of accuracy demonstrates the wide distance between two levels of human thinking.

Problems of tangents and arcs. The major part of the mathematical work of Archimedes is related to what we now call Calculus. For example, the determination of the *tangents* to the "Archimedean" spirals is related to the differential calculus. In the work preserved by Eutokius (500 A.D.) we find the solution of the problem of determining the maximum of the function $x(s-x)^2$. This is done by determining a hyperbola of the type $xy = c$ which touches the parabola $y = (s-x)^2$. We see that Archimedes solved the problem of finding the maximum through the use of tangents as we do it now.

Much more important are the problems connected with the process of *integration*. We have already mentioned the determination of the area of the surface of a sphere with given radius, and the determination of its volume. Archimedes started with the computation of the area of the surface and, for this purpose, had to compute the integral of $\sin x$. He achieved this integration by establishing the identity

$$\sin \frac{\pi}{2n} + \sin \frac{2\pi}{2n} + \sin \frac{3\pi}{2n} + \cdots + \sin \frac{(2n-1)\pi}{2n} = \cot \frac{\pi}{4n}.$$

Characteristic for the trend of Greek mathematics is the formulation of this problem: it is required to construct a plane figure having the same area as the surface of a sphere with given radius. The answer is: the plane figure is a circle with radius twice that of the given sphere. Having found the area of the surface, Archimedes determined the volume.

Further we find Archimedes dealing with the construction of a square with area equal to that of a plane figure bounded by arcs of a parabola and by straight lines. For the squaring of the parabola, he had to determine the asymptotic value of the sum

$$\frac{1^2 + 2^2 + \cdots + n^2}{n^3}.$$

In another solution of the problem, Archimedes makes use of mechanical notions, primarily of the center of gravity and its obvious properties. (Incidentally, it is not easy to take the existence of the

center of gravity of an arbitrary body for granted.) He uses these properties to evaluate, as we shall say, an integral.

This is also the method he uses in a work written for Eratosthenes, an outstanding astronomer and mathematician of his time. This work was found only recently in Istanbul, written on parchment that was later used for a liturgical text. Here Archimedes combines the mechanical method with the device of taking areas or volumes as if these were aggregates of lines or surfaces. Everywhere else in his works he uses the exact method of exhaustion. In this latter one, dedicated to a fellow master of mathematics, he does not hesitate to bring into play the relations between properties of areas and properties of a sum of lines. In this work he finds the *volume* of a sphere directly (thereby avoiding the integration of the sine function), reducing the problem to that of the measurement of the cone.

There are probably extant in Arabic translation some other works of Archimedes. Not long ago there was discovered in the work of an Arabic author a remarkable construction of the regular heptagon by Archimedes (see Figure 2). It is as follows: $ABCD$ is a square. A line $A E F G$ is so constructed that it meets the diagonal DB in E , the side BC in F , the side DC in G , so that the triangles $A E B$ and $F C G$ have the same area. H is a point on DC such that $E H$ is parallel to $C B$. Then $H C$ is the side of the regular polygon with 14 sides in the circle with the radius $C G$.

Such constructions with a moving ruler meeting two given lines in a prescribed way were often used in Greek mathematics.

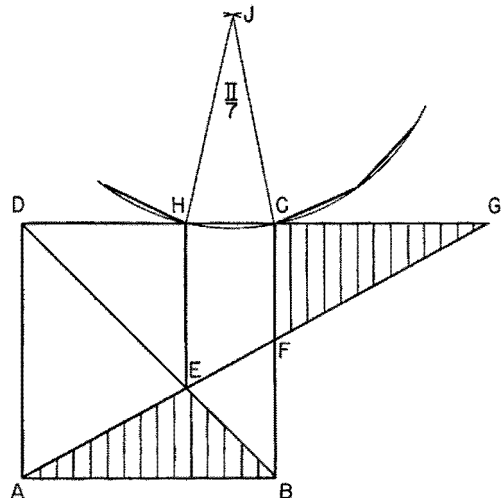


Figure 2.

3.3 The achievements of Apollonius

Apollonius, some forty years younger than Archimedes, was born in Pergamum, about 250 B.C. Pergamum was, at this time, beginning to be a center of culture. There are still extant great works of art produced there at the time of Apollonius. It is probable that he lived a part of his life in Alexandria, but he was closely connected with at least one of the rulers of Pergamum, Attalus, to whom he dedicated one book of his great *Treatise on conics*.

Whereas Archimedes made his investigations in regions hitherto untrodden and used new methods, Apollonius uncovered in his principal work, the treatise on conics, a field of geometry already partially known as a grand structure of admirable wealth and beauty. Through the theory of conics the astonishing fecundity of mathematical thought became apparent, perhaps for the first time. One simple notion, in this particular case, that of the plane sections of a cone, produces a wealth of problems and theorems.

One aim of Apollonius was to cover in a systematic treatment all that was already known about conics. Originally, the conics were considered as spatial phenomena. But we find Archimedes already using a definition of the conics as plane loci. The points of these loci are defined by a pair of lines which are different from our common rectangular coordinates only in so far as they are not a pair of proportions (or numbers) obtained through the introduction of unit lines. In Archimedes the “equation” of the conic has the form

$$y^2 = kx(d - x),$$

where k is a proportion (or a number) (see Figure 3). For $k > 0$ and $k \neq 1$ we get the ellipse immediately as the projection of the circle for which $k = 1$. In Apollonius, the equation for the conics, the basis for all his investigations, is in oblique coordinates

$$y^2 = px \pm \frac{px^2}{a},$$

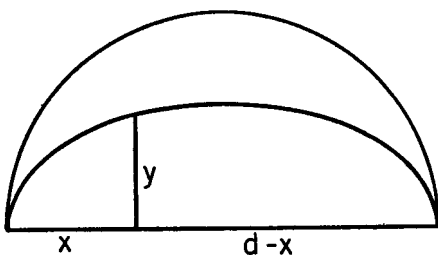


Figure 3.

where the + sign gives the hyperbola (y^2 is “surpassing” px), the – sign characterizes the ellipse (y^2 is “deficient” with regard to px); for infinitely great a we get the parabola (y^2 “equals” px). These names for the different conics are probably Apollonius’ own.

Apollonius had to solve many problems to achieve a systematic theory of the conics. In the original equation for the conic, there are three parameters: p , a , and the angle between the x -line and the y -line. From these data one had to find the position and the length of the axes. Further, to have theorems valid for both ellipse and hyperbola, it was necessary to take the two branches of the hyperbola together as one curve, a difficult abstraction since it seems to contradict the appearances. This abstraction became easy only after the introduction of infinitely distant points in the seventeenth century.

In the work of Apollonius there are found the elementary construction of tangents to a given conic through a given point or in a given direction, the discussion of the problems of normals through a given point, and geometric constructions carried out with the help of conics. These problems are related to problems of maxima and minima and to the discussion of the intersection of conics. This latter problem is equivalent to the determination of the number of real roots for an equation of the fourth degree by means of certain inequalities between the coefficients.

The problem of determining a conic by five of its points is not considered, in spite of the solution of equivalent problems. Probably the obstacle was that the problem could not be easily enunciated, because it was necessary to determine at the same time whether one could construct an ellipse, hyperbola, or parabola going through the five given points (the assumption being that no three of them are collinear).

It seems quite impossible to find out who it was that determined the foci. One wonders in what way the old mathematicians came to discover these characteristic points. Probably Euclid already knew that the conics were loci of points for which the distances from a fixed point and from a fixed line have a fixed ratio. Apollonius knew the most important properties of the foci of the ellipse and hyperbola, namely, that they are the centers of the orthogonal involutions, that the tangents make equal angles with the lines through the foci, and finally the theorem concerning the sum or difference of the focal distances.

We conclude our report on Archimedes and Apollonius with a remark by Leibnitz: He who under-

stands Archimedes and Apollonius will admire less the achievements of the foremost men of later times.

4 200 B.C.–600 A.D.

4.1 Trigonometry

It is in this period that we find Greek mathematics strongly influenced for the first time by phenomena outside the world of mathematical ideal entities. The astronomer observes the positions of the sun, moon, planets and fixed stars at different places and at different times. He measures a number of angles which are not independent variables. To establish their relations, to determine other, not observable, angles, for instance the angle measuring the arc between the annual circular path of the sun and the pole of the daily circles of the fixed stars, requires new mathematical investigations. All these problems pertain to the geometry of the sphere and the solution of these problems constitute what is now called spherical trigonometry.

Furthermore, we find plane trigonometry developed. The investigation of relations between the sides and the angles of a plane triangle was perhaps inspired by spherical trigonometry.

Now, in the case of spherical trigonometry, as in that of plane trigonometry, the relations are not algebraic if one measures the angles as fractional parts of a full angle, which is always the case in astronomical observations. One needs to introduce non-algebraic functions of the fractions determining the angles in order to obtain algebraic relations. It is sufficient to introduce one function of this kind: One may take the angle as an angle between the radii of a circle, then this function can be chosen as the ratio of the chord subtending the angle to the diameter of the circle.

Thus we have two problems for spherical and plane trigonometry:

1. One has to find the (algebraic) relations between the chord function of the angles and the arcs in a spherical triangle and the (algebraic) relations between the chord function of the angles and the sides of a plane triangle.
2. One has to investigate the chord function, which again may be divided into three parts:
 - (a) the determination of the algebraic functional relations;
 - (b) the numerical computations of the function, which in turn implies

(c) the finding of certain inequalities determining the behavior of the function in the neighborhood of certain points.

We note in passing that the chord function of an angle is double the sine function of the half of the angle.

4.2 The seeds of the notion of function and of transformation

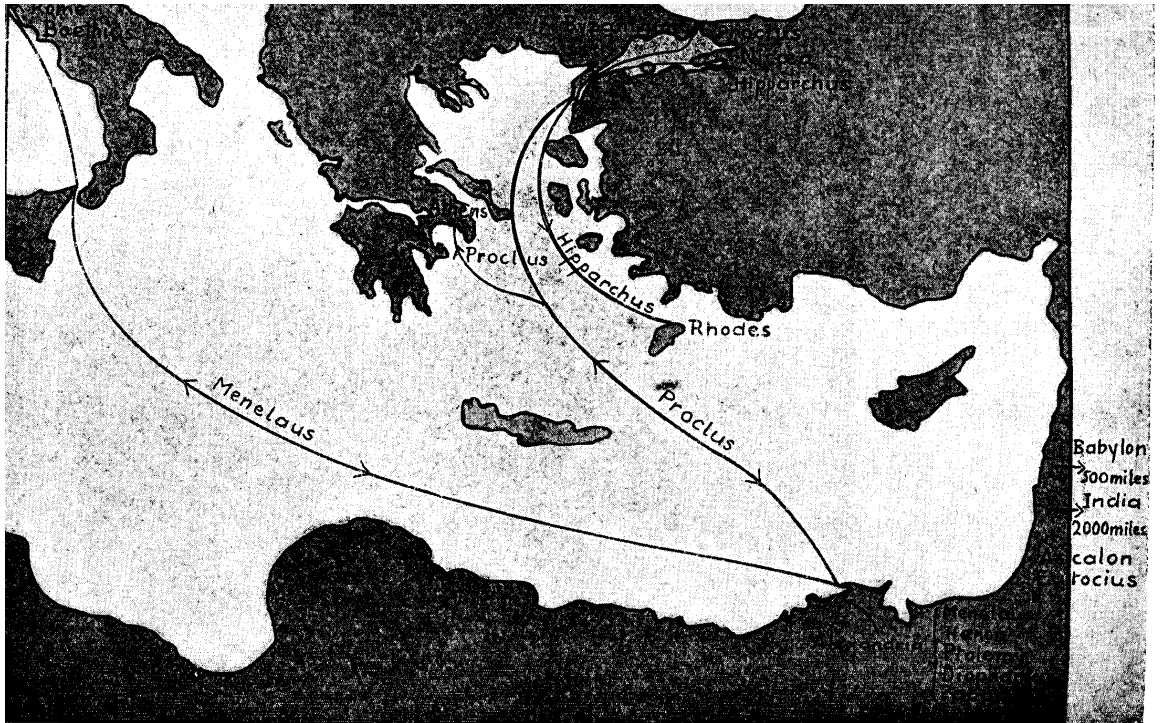
In trigonometry the mathematicians came closer to the notion of function than in the older theory of “locus”. But the general idea of function did not appear at all, still less the idea of transformation. A transformation is, one may say, nothing else than a materialized function: the transformed geometrical element, for instance a new point, is a function of the old one. But this embodiment of a function is not so easily visualized as the locus. We shall see below that the main difficulties confronting the foundation of projective geometry are overcome in this period. But still we do not find the beginning of a systematic development of projective geometry. We do not even find a systematic use of the fact that the conics are generated by the projective transformation of a circle.

4.3 Commentaries

In this period we find several valuable commentaries on the works of the mathematicians Euclid, Archimedes, and Apollonius.

4.4 Other peoples

The disturbances of the old world by the expeditions of Alexander had also an effect on the state of mathematics. The Greeks came into closer contact with the Egyptians and with the people of Mesopotamia. The inhabitants of India came in contact with Greek art and Greek mathematics, also in closer contact with the culture of Asia Minor. Although Greek mathematics during its first period was certainly under Babylonian influence, its peculiar and vigorous development obscured that influence for many centuries. At this time, however, the characteristics of Babylonian mathematics became quite apparent. Mathematical exercises, to be seen in cuneiform texts from 2000 B.C., appear in Greek textbooks. In India, we see a flourishing of mathematics under Greek influence, especially arithmetic which, perhaps, was reflected back to the Greeks. For the first



time after Euclid we find in this period new arithmetical problems and new methods for their solution. These were again seeds which, 1500 years later, were developed into full growth.

Important new symbols for numbers were introduced in India and were transmitted to the Near East and Europe where we shall encounter them in the next period.

4.5 General significance of this period

The time covered by this article is very long, 800 years, in comparison with the periods covered by earlier sections. In it no mathematician of such glorious fame as Euclid, Archimedes, or Apollonius appears. But it is not easy to call it a period of decadence if one recognizes the many seeds to be developed later. One mathematician of this time, Pappus (about 300 A.D.), even expressed his feeling that mathematics up to his time was only in the beginning of its development. He says: "I saw that all (mathematicians) move about only in the beginnings of pure and applied mathematics; and I had a feeling of awe (because I was aware) that I could show much better and much more useful things." Perhaps Pappus was afraid to enter alone into the vast and unknown realm of that science the existence of which he divined. Already in Pappus' time the best minds were

more interested in mystical or theological problems than in scientific ones.

The great treatise of Ptolemy of Alexandria (about 150 A.D.) on astronomy stood for more than 1500 years by the side of Euclid's *Elements* as a book of indisputable authority.

4.6 The foundations of trigonometry

This treatise of Ptolemy is commonly called *Almagest* into which word the Arabs changed the original title, *ἡ μεγάλη σύνταξις*, *The Great Composition*. In this book Ptolemy collected, enlarged and systematized the results of preceding astronomical investigations, both practical and theoretical, especially those of the great Greek astronomer Hipparchus (about 200 B.C.). Most famous is his systematization of the movement of the planets by using combinations of circular movements. But this part of his work is of minor interest for the history of mathematics.

The trigonometry of the *Almagest* is based on two theorems. The first is called the theorem of Menelaus, who lived about fifty years before Ptolemy in Alexandria. His work is extant only in Arabic and Hebrew translations. Figure 4 represents the theorem of Menelaus for plane figures, where we

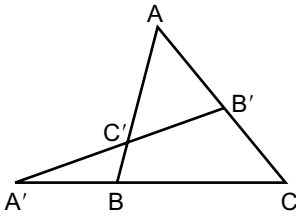


Figure 4.

have the relation

$$\frac{A'B}{A'C} \cdot \frac{B'C}{B'A} \cdot \frac{C'A}{C'B} = 1.$$

This theorem is also valid for the sine function of the arcs on the sides of a spherical triangle and in this form is used by Ptolemy to prove the relations between the arcs and angles of a spherical triangle. It is interesting to see why the theorem for the plane figures is so easily changed into a theorem for spherical figures: the theorem in the plane is easy to generalize for the plane of projective geometry by taking cross ratios instead of ratios. Then we obtain immediately the corresponding theorem about planes and lines through a point which we take as center of the sphere. The cross ratio of points on a line in the plane corresponds to the cross ratio of lines in a plane through the central point. And this latter cross ratio is expressed by the sine function of the angles between the lines; hence the advantage of our sine function in comparison with the original chord function. Menelaus, of course, as well as Ptolemy, does not go the way of projective geometry to prove the spherical theorem by the plane theorem.

The second theorem is used to find the functional relations for the chord function. This theorem is called after Ptolemy and states a relation between the six distances of four points on a circle. The proof of Ptolemy is probably the shortest possible, and is to be found in all textbooks on geometry. The analysis of the theorem shows the following elements: first, an identity between two cross ratios, determined by the same four points, an identity known as Euler's identity for four points on a line; second, the algebraic identity expressing the invariance of the cross ratio under a linear transformation; third, the geometric fact that a linear transformation in the plane of a complex variable transforms lines into circles. The theorem is used in the special case where two of the four points are on one diameter. To go, by way of this theorem, to the addition theorem for the sine function is certainly not the simplest possible way.

For the numerical evaluation of the chord function Ptolemy needs the addition theorem and further the fact that the function $(\sin x)/x$ is decreasing with increasing x for $0 < x < \pi/2$. The fact that $(\tan x)/x$ is increasing with increasing x for $0 < x < \pi/2$ is already proved in a very simple way in Euclid's *Optics*. One may prove in a similar and quite as simple way that $(\sin x)/x$ is decreasing with increasing x for the same range. Ptolemy gives a rather complicated proof. The fact itself had already been used by Aristarchus of Samos more than three hundred years before Ptolemy when he discussed the appearance of the sphere of the moon. Tables for the chord function were given long before Ptolemy by Hipparchus but these have not been handed down to us.

4.7 Achievements in geometry

The most remarkable achievements of this period in geometry are due to Pappus. In the seventh book of his *Mathematical Collections* he proved the theorem that the cross ratio of four points on a line $(AC/AD):(BC/BD)$ is not altered by perspective projection. With the help of this theorem he proves several other propositions. By far the most important is the following: Let A, B, C be three points on one line, A', B', C' three points on another line; then the lines AB' and BA' , BC' and CB' , CA' and AC' , respectively, meet in three points lying on one line.

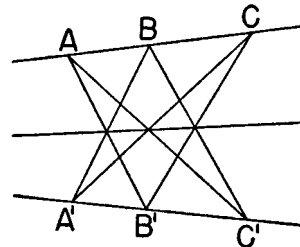


Figure 5.

This theorem marks an event in the history of geometry. From the beginning geometry was concerned with measures: lengths of lines, areas of plane figures, volumes of bodies. Here we have for the first time a theorem which is established by the ordinary theory of measures but is itself free of all elements of measurement; it states the existence of a figure which is determined through the incidence of lines and points only. It is the first "configuration" of projective geometry, and it was shown more than 1500 years later that this configuration alone is sufficient to build up projective geometry in the plane. The

mathematicians who pointed out the important role of this theorem were unjust toward Pappus in naming the theorem after Pascal.

4.8 Practical mathematics

After Euclid's *Elements* and Ptolemy's *Almagest* there is perhaps no ancient work on mathematical methods and natural science which had such a lasting, uninterrupted influence as that of Heron. About Heron's life we know scarcely anything. For a long time scholars tried to find out when he lived through a study of his works, and made guesses ranging from 150 B.C. to 200 A.D. For the moment it seems to them most probable that he lived in Alexandria in the first century A.D. There are books bearing his name which probably have not been written by him. In this whole collection we find little pure mathematics. It is only incidentally that, teaching all sorts of practical methods of measuring, he states and proves rigorously how to express the area of a triangle in a symmetrical way in terms of the sides. He also shows himself as a resourceful mathematician in the proof of the theorem that the three auxiliary lines in Euclid's proof of the Pythagorean theorem meet in one point. But the main tendency of Heron is to make mathematics bear fruit in the treatment of practical problems. He shows many ways of finding approximate measures.

He also treats problems of surveying, for instance the problem of finding the direction of the line joining two points, if one of the points is not visible from the other point. He solved this problem by measuring the rectangular coordinates of auxiliary points between the two given points. This is in symbols: x_0, y_0 and x_n, y_n may be the coordinates of the two given points, x_i, y_i ($i = 1, \dots, n-1$) the coordinates of the auxiliary points; then the direction is given by

$$\frac{y_n - y_0}{x_n - x_0} = \frac{\sum_{i=1}^n (y_i - y_{i-1})}{\sum_{i=1}^n (x_i - x_{i-1})}.$$

The idea of determining the position of different points on a surface, especially the earth, by two coordinates is much older than its appearance in Heron's work. Coordinates were quite indispensable to the astronomer in determining the relative positions of the places of observation. Thus we find already Hipparchus determining the points on the globe through the two angles of latitude and longitude.

In many of the problems of Heron concerned with measuring we find similarities with Egyptian meth-

ods. There are other aspects of his work which are obviously connected with the mathematics of the Babylonians. Thus we find an example of a quadratic equation for the radius of a circle given through the sum of the number measuring the area of the circle and of the number measuring the circumference. Such problems as this have scarcely any practical value; furthermore, they are very remote from the problems of classical Greek mathematics. Similar problems, however, are to be found in very old collections of Babylonian exercises.

Heron also tries to compute the volume of a truncated pyramid whose linear measures as they are given are impossible. Out of these data he gets for the volume an expression corresponding to the square root of a negative number. He takes instead of this expression the square root of the number with positive sign, which magnitude, within the given problem, has no significance whatsoever.

We may say that there are in Heron's works characteristics of the Greek, Egyptian, and Babylonian mathematics, and even a premonition of developments of much later times.

Heron was not only an ingenious mathematician. He observed the forces of nature and used them to build all sorts of machines, most of them of very little practical use. He was probably the first to use the power of expanding steam to set heavy bodies in motion, for instance to make a sphere rotate about an axis. His manifold inventions were well known to the Renaissance physicists, among others, to Galileo.

4.9 Algebra and theory of numbers

There is a fourth mathematician in this period whose name still lives in the work of modern mathematicians, Diophantus. In the first book of his *Arithmetic* we find many problems not very different from old Babylonian problems. But the form of his treatment of these problems is very important: as the problems never lead to an irrational quantity, Diophantus is able to present a theory of equations in a seemingly modern form. All the difficulties in operations with irrational quantities, which can only be overcome by an at least partially developed theory of limits, do not appear here. Thus we find here symbols and methods of solution of equations quite similar or equivalent to modern symbols and methods. The form of Diophantus' work has undoubtedly influenced the further development of algebra. However, he was probably not the first to give this form to arithmetical operations.

We know nothing of the life of Diophantus. Those scholars may be right who suppose that Diophantus lived at Alexandria in the time of Ptolemy and Heron.

Beginning with the second book of Diophantus' *Arithmetic* we find a new type of problem, belonging to the theory of numbers. In the first problem of this type one has to find two rational numbers which squared and added are equal to a given square, a^2 . The method of the solution is quite general. He puts one number equal to x , the other equal to $rx - a$. Then he finds

$$x = \frac{2ar}{r^2 + 1}.$$

He indicates the general solution but takes special values for a and r . This problem is, of course, not different from the old problem of finding two rational numbers x and y so that $x^2 + y^2 = 1$. But already the next problem is something new: to divide the sum of two squares $a^2 + b^2$ into two other squares. He puts one number equal to $x + a$ and the other number equal to $rx - b$ and finds

$$x = \frac{2rb - 2a}{r^2 + 1}.$$

Again he indicates a general solution but takes special values for a , b , and the parameter r . This procedure is nothing else than the rationalization of the equation $u^2 + v^2 = a^2 + b^2$.

In the same way the next example is to be considered as a rationalization of the algebraic relation $y^2 - z^2 = d$. In general we may characterize the problems of Diophantus as problems of rationalization of algebraic relations, i.e., of the finding of a representation of an algebraic relation through rational functions of a parameter. Diophantus does not try to find solutions in integers.

These problems and their solutions made a great impression on mathematicians of the sixteenth and the seventeenth centuries and were certainly one of the reasons for a new flowering of that noble science, the theory of numbers.

4.10 Commentaries

The most famous of the ancient commentators is doubtless Proclus, who lived in the middle of the fifth century A.D. But he was much more a philosopher than a mathematician. He was head of the Neo-Platonic school at Athens. His contributions to mathematics are certainly very slight, but his commentary on the first book of Euclid is an invaluable source for

the history of Greek mathematics. The commentary is also typical of his time, which considered metaphysical speculations, mostly of a mystical character, as the most important task for lovers of wisdom. These people looked down on mathematics because it made use of hypotheses whereas pure speculation was nonhypothetical.

Heron, too, made additions to, and commentaries on, the work of Euclid. Pappus wrote about Euclid, Apollonius and other mathematicians. Of the later commentators we mention Eutocius, who lived in the sixth century A.D. and came from Ascalon in Syria. He showed himself a very able mathematician and gave us an excellent commentary on Archimedes, where we find the most valuable report on the different solutions of problems of the third degree. He was even able to restore an old corrupted manuscript, probably of Archimedes, where we find the solution of a maximum problem, mentioned in our third section.

Here we make an end to our necessarily incomplete report on Greek mathematics.

4.11 The Romans

We mention only one Roman, Boethius, who was executed by Theodoric, King of the Goths, in 524. His mathematical importance lies in his role as translator. In his time Greek was no longer known by all who were interested in science. Boethius translated, along with Plato and Aristotle, also Euclid, Ptolemy, and Archimedes, but these translations are not extant. He is the first author where we find the quadruple of the four sciences, arithmetic, music, geometry, and astronomy, as constituting the mathematical branch of the liberal arts, the quadrivium. We do not owe to the old Romans any significant contribution to mathematical science.

4.12 Mathematics in India

We can give only a very short review of mathematical activities in India during this period. The outstanding mathematicians of this country were primarily astronomers; we mention only Aryabhata (about 500 A.D.) and Brahmagupta (about 600 A.D.). Whereas Diophantus treated problems of finding rational values for algebraic relations of second and higher degree, the mathematicians of India treated the problem of finding integers satisfying linear relations. One recognizes that such problems are indeed of interest in astronomical investigations. It is prob-

able that the Indians were depending on the research work of Babylonian astronomers.

Probably the Indians, in their development of a new symbolism for the writing of integers, were also indebted to the Babylonians. They determined the integer by expanding it into a power series with 10 as basic number:

$$n = \sum_{i=0}^m a_i 10^i, \quad a_i < 10.$$

Then they represented the integer by the symbol

$a_m \cdots a_2 a_1 a_0$. The old Babylonians had already represented integers in this way, using 60 as the basic number. But their representation was ambiguous because they did not use symbols for those coefficients which are equal to zero.

We find our sign 0 already in the astronomical tables of Ptolemy, and we also find it used as an abbreviation for “nothing” in Heron’s writings. But the systematic use of the symbol 0 came from India to the Arabs and through them to Europe, and had an inestimable influence upon all kinds of scientific and practical computations.

Diophantus of Alexandria

J. D. SWIFT

American Mathematical Monthly 63 (1956), 163–170

1 Introduction

The name of Diophantus of Alexandria is immortalized in the designation of indeterminate equations and the theory of approximation. As is perhaps more often the rule than the exception in such cases, the attribution of the name may readily be questioned. Diophantus certainly did not invent indeterminate equations. Pythagoras was credited with the solution

$$(2n + 1, 2n^2 + 2n, 2n^2 + 2n + 1)$$

of the equation $x^2 + y^2 = z^2$; the famous Cattle Problem of Archimedes is far more difficult than anything in Diophantus, and a large number of other ancient indeterminate problems are known. Further, Diophantus did not even consider the most common type of problem called by his name, the linear equation or system of equations to be solved in integers.

Nevertheless, on at least three grounds the place of Diophantus in the development of mathematics is secure. On all the available data he was the first to introduce systematic algebraic procedures to the solution of non-linear indeterminate equations and the first to introduce extensive and consistent algebraic notation representing a tremendous improvement over the purely verbal styles of his predecessors (and many successors). Finally, the rediscovery of the book through Byzantine sources greatly aided the renaissance of mathematics in western Europe and stimulated many mathematicians, of whom the greatest was Fermat. (Much of Fermat's work is known from notes written in his copy of Diophantus [1].)

Of Diophantus as an individual we have essentially no information. A famous problem in the *Greek Anthology* indicates that he died at the age of 84, but in what year or even in which century we

have no definite knowledge. He quotes Hypsicles and is quoted by Theon, the father of Hypatia. Now Hypsicles, in the introduction to his book, the so-called *Book XIV* of Euclid, places himself within a generation or so of Apollonius of Perga whose time is definitely established by the rulers to whom he dedicates his works. Thus we may put Hypsicles in the early or middle part of the second century B.C. with reasonable accuracy [17]. Theon, on the other hand, definitely saw the eclipse of 364 A.D. [10]. Within this gap of five hundred years, historians are at liberty to place Diophantus wherever he best fits their theories of historical development [10, 14]. The majority follow [2] and, on the basis of a dubious reference by the Byzantine Psellus (c. 1050), assign him to the third century A.D.

2 The *Arithmetic*

The surviving work of Diophantus consists of six books (sometimes divided into seven) of the *Arithmetic* and a fragment of a work on polygonal numbers. The introduction to the *Arithmetic* promises thirteen books. The position and content of the missing six or seven books is a matter of conjecture. (The reader is reminded that a "book" is a single scroll and represents the material contained in twenty to fifty pages of ordinary type.)

These books may be summarized as follows: *Book I*: Determinate systems of equations involving linear or quadratic methods. *Books II to V*: Equations and systems of equations, the majority of which are quadratic indeterminate although *Books IV* and *V* contain a selection of cubic equations, determinate and indeterminate. *Book VI*: Equations involving right triangles. All books consist of individual problems and their solutions in positive rationals. In

the ordering of the problems some consideration has been given to relative difficulty and interrelation of material, but the over-all impression is of a disconnected assortment.

3 Notation

The numerical notation used by Diophantus is, of course, the standard Hellenistic notation which uses the letters of the Greek alphabet with three archaic letters added to give 27 different symbols [6, 7]; the first nine stand for units, the second for tens and the last for hundreds. Thus, for any further notation, either non-alphabetic symbols or monogrammatic characters were required.

There is a single symbol for the unknown quantity. This may be a monogram for $\alpha\rho\iota\theta\mu\omicron\varsigma$. The symbol for “minus” is apparently a monogram for the root of $\lambda\epsilon\iota\psi\iota\varsigma$ [5]. Addition is indicated by juxtaposition. The powers of the unknown are designated by easily recognizable monograms, the square by Δ^v for $\delta\upsilon\nu\alpha\mu\iota\varsigma$, the cube by K^v for $\kappa\nu\beta\omicron\varsigma$. Higher powers are formed from these by addition, i.e., the fifth power is considered as square-cube. To avoid ambiguity it is necessary to have a special symbol for the zero-order terms also, a monogram for $\mu\omicron\nu\alpha\delta\omicron\varsigma$, and to write all the negative terms together. Thus, if we adopt an equivalent set of conventions in English, retaining Arabic numerals and the letter x for the indeterminate, the expression

$$6x^4 + 23x^3 - 2x^2 + x - 5$$

would appear as

$$S^q S^q 6C^u 23X1MS^q 2U^n 5.$$

Fractions were represented either in the inverse position to the present day or by inserting the word for “divided by” between the numerical expressions on the same line. Reciprocals of integers and negative powers of the unknown are designated by a special symbol placed after the number or power.

The most important limitation of this notation is the restriction to one unknown. Since practically all the problems require the determination of several quantities, a considerable part of Diophantus’ work lies in the reduction to a single quantity. Further, no general solution in expressed parameters is possible. Even if a general method is indicated, it must be restricted in its presentation to a specific numerical case.

A particular problem will illustrate the situation. In problem 1, Book IV, it is desired, in modern terms, to solve the system:

$$x^3 + y^3 = a, \quad x + y = b.$$

Essentially the method is to let

$$x = z + b/2, \quad y = b/2 - z.$$

Substitution in the first equation now yields a binomial quadratic. Let us look at this problem in a translation as bald as possible:

To partition a given number into two cubes of which the sum of the sides is given: Let the number to be partitioned be 370 and the sum of the sides $U^n 10$. Let the side of the first cube be $x1U^n 5$, the latter term of which is half the sum of the sides. Therefore, subtracting, the side of the other cube is $U^n 5Mx$. Then the sum of the cubes will be $S^q 30U^n 250$. This is equal to $U^n 370$ as is given and x becomes $U^n 2$. As to the original numbers, the first side will be $U^n 7$ and the second, $U^n 3$. The first cube, 343; the second, 27.

4 Diophantine algebra

With this problem in mind let us turn to some aspects of Greek and Babylonian mathematics. A number of tablets [15, 16], both old Babylonian (1800–1600 B.C.) and Seleucid (300 B.C. and later), exist which teach the solution of equations which can be reduced to the forms

$$x + y = a; \quad xy = b$$

[10, 12, 13]. Again Euclid’s *Elements* II, 5, 6 can best be viewed as giving solutions to these problems [12]. In modern notation, the procedure in both cases is to write

$$x = a/2 + z, \quad y = \pm(a/2 - z);$$

$$xy = b = \pm(a^2/4 - z^2); \quad z = \sqrt{a^2/4 \pm b}.$$

Now Diophantus in I (27, 30) considers the same equations, solves them the same way and applies the basic idea repeatedly as in the quoted problem. Other examples can be followed in a similar way, e.g.,

$$x^2 + y^2 = a, \quad xy = b.$$

(See [13] for a complete discussion of quadratic equations in antiquity.)

Let us now compare the treatment in the three cases. The tablets consist of lists of problems of varying complexity each framed in specific numbers and quantities. The problems are not “practical” nor in any sense rigorously geometrical; men are added to days, lengths to areas, areas are multiplied, etc. It is clear that the basic thought is purely algebraic. The problems are so set that the solutions are positive integers or terminating sexagesimal fractions such that the roots can be obtained from tables of squares, but from other tablets we learn of approximations to non-terminating rationals like $1/7$ or irrationals like $\sqrt{2}$.

The Euclidean problems are cast in the form of propositions about line segments, squares and rectangles. Their generalizations in II, 28, 29, concern parallelograms. The propositions are general and the result is deduced rigorously from the postulational basis. The results are line segments which may well be incommensurable with the original segments; i.e., “irrational” answers are acceptable.

In Diophantus the problems are formulated in terms of abstract numbers but a “number” is always positive rational. The solutions are worked out in terms of particular numerical examples. This procedure may be considered analogous to carrying out a geometrical construction in terms of particular line segments and, indeed, Diophantus probably intended that his problems should be read in this manner. There is, however, no pretense at postulational development. No general propositions are stated even where the solution implies them. Restrictions on the choice of initial values are not always given; in the case of I, 27, we are informed that $a^2/4 - b$ must be a square but in the problem in the previous section no restriction is mentioned. The most reasonable conclusion is that he did not know the form of the restriction or did not know how to express numbers that satisfied the restriction. The authors of [5] and [4] disagree but on the naive ground that since he did come up with workable numbers 370 and 10, he must have had some way of generating them. The answer is obvious; he generated them from the answer.

Like the Babylonians, Diophantus had no qualms about adding areas and lengths (see VI, 19 in [5]) although, to be precise, he says that he adds “the number in the area” to “the number in the length”. His algebraic technique is tremendously advanced beyond anything we possess of the Babylonians. The complicated cubic and higher degree equations and the indefinite equations are not even suggested in

Babylonian algebra. The latter had examples of binary cubics and a few other higher degree equations soluble by tables; they also knew general forms for Pythagorean numbers and obtained solutions of $x^2 - 2y^2 = 1$, but this is as far as our present evidence takes them. Even in the quadratic case there may be a difference [13]. When a quadratic is to be solved, Diophantus makes some effort to choose the variable so that a binomial equation results, but if this is not practicable, the general quadratic formula (positive sign before the radical) is used without further comment. The question is still at issue whether the Babylonians ever solve a quadratic without bringing it into some normal form involving a known sum or difference and product.

It is useless even to try to guess what proportion of the advanced problems and methods are Diophantus’ own. Most modern historians postulate a continuous underlying tradition of oriental algebraic methods in Greek mathematics rather than a sudden invasion in the Roman period. If this be so, texts and problem lists would certainly have existed. It is probable that the *Arithmetic* was in good part a compilation of such a quality that the predecessors were no longer held in repute. There are traces of the Diophantine notation elsewhere; Heron (60 A.D.) used the same minus sign for example, but no evidence exists that the semi-algebraic notation or the general methods it permitted were used before the publication of the *Arithmetic*.

To sum up, the basic algebraic approach in Diophantus is Babylonian. The generality and abstraction is Greek. The work may be viewed as an episode in the decline of Greek mathematics [12] or as the finest flowering of Babylonian algebra [10].

5 Indeterminate problems

In giving translations of several illustrative problems, I have avoided the usual practice of direct translation. Instead, I adhere carefully to the method of the original while replacing the particular numbers used by parameters. The rationale may thus be conveyed with less verbal explanation than if the presentation were given in its original special form. At the same time, the full power of the method is apparent.

II. 9: If $n = a^2 + b^2$, find other representations of n as the sum of two squares.

Modify a to $x+a$. The corresponding modification of b may be written $(rx-b)$. Here x is the unknown,

a , b and r were assigned specific values.

$$\begin{aligned} n &= a^2 + b^2 = (x + a)^2 + (rx - b)^2 \\ &= (r^2 + 1)x^2 + (2a - 2br)x + a^2 + b^2. \end{aligned}$$

Thus, $x = (2br - 2a)/(r^2 + 1)$, where r may be any rational such that the required quantities are positive.

Note the clever choice of the unknown; fixing x and solving for r would leave a condition on x still to be met; $b^2 - x^2 - 2ax$ would have to be made a square. Again, if a is increased by a fixed amount and the unknown is taken as the corresponding decrease in b , the result does not come out at once. The choice of $rx - b$ instead of $b - rx$ was dictated solely by the numerical values selected which happened to make $b - rx$ negative. (But see *V*, 24, below.) Euler wrote

$$a^2 + b^2 = (a + rx)^2 + (b - qx)^2$$

which results in a more symmetric solution but this concept is foreign to Diophantus' notation and the solution above is quite general.

Here would be a perfect opportunity to state a proposition instead of a problem. A proof of the theorem: "Any number which is the sum of two squares can be represented as such in an infinite number of ways", is contained in the solution above.

III. 6: Find three numbers whose sum is a square and such that the sum of any two is a square:

$$\begin{aligned} x + y + z &= t^2 \\ x + y &= u^2, \quad y + z = v^2, \quad x + z = w^2. \end{aligned}$$

Here Diophantus assigns a definite value to w , or, in modern notation, lets it play the role of the parameter. He then chooses an unknown, r , restricting it as follows: Let $t = r + 1$, $u = r$, $v = r - 1$. Then $z = 2r + 1$, $y = r^2 - 4r$, $x = 4r$ and $w^2 = 6r + 1$. Thus $r = (w^2 - 1)/6$ where w is an arbitrary rational exceeding 5 (so that y is positive). So

$$\begin{aligned} x &= (2w^2 - 2)/3; \\ y &= (w^2 - 1)(w^2 - 25)/36; \\ z &= (w^2 + 2)/3. \end{aligned}$$

This problem was chosen to illustrate two points. First, Diophantus is not interested in generality except as an incidental by-product. A considerable increase in generality can be obtained merely by replacing $r \pm 1$ by $r \pm s$ in the solution and this possibility could easily have been indicated by the addition of a single phrase. Second, the choice of wording of the problems is often peculiar from a modern

viewpoint. This problem is clearly equivalent to the single equation:

$$u^2 + v^2 + w^2 = 2t^2.$$

Incidentally, using methods available to Diophantus but probably exceeding his control of notation, a much more general solution of this equation is available than the system

$$(u, v, w, t) = ((w^2 - 1)/6, (w^2 - 7)/6, w, (w^2 + 5)/6)$$

given above. The equation being homogeneous, it will be more convenient to solve in integers. Let $w = rs$, $u = s^2 - p$, $v = s^2 - q$, then

$$s^4 + (r^2/2 - p - q)s^2 + (p^2 + q^2)/2 = t^2.$$

The left-hand side is a perfect square if

$$p^2 + q^2 = 2k^2, \quad r^2/2 - p - q = 2k.$$

The first of these is the problem of finding three squares in arithmetic progression. It does not occur specifically in the *Arithmetic*, probably because it is too simple in the rational case, reducing essentially to $a = b = 1$ in the problem above. It will be more convenient to take a solution derived from the solution given to the Pythagorean equation by Euclid. If $X^2 + Y^2 = Z^2$,

$$(X + Y)^2 + (X - Y)^2 = 2Z^2.$$

Thus

$$\begin{aligned} p &= -m^2 + 2mn + n^2, \quad q = m^2 + 2mn - n^2, \\ k &= m^2 + n^2, \quad r^2 = 4(m + n)^2, \\ &\text{and } r = 2(m + n). \end{aligned}$$

Thus

$$\begin{aligned} (u, v, w, t) &= (s^2 + m^2 - 2mn - n^2, \\ & \quad s^2 - m^2 - 2mn + n^2, \\ & \quad 2(m + n)s, s^2 + m^2 + n^2). \end{aligned}$$

The previous solution is obtained by setting $m = 2$, $n = 1$ and dividing by 6.

V. 24: Find a solution of $x^4 + y^4 + z^4 = t^2$.

If $t^2 = (x^2 - m)^2$, then $x^2 = (m^2 - y^4 - z^4)/2m$. Thus an integer m must be found so that

$$(m^2 - y^4 - z^4)/2m$$

is a square. Let $m = y^2 + z^2$ so

$$x^2 = y^2 z^2 / (y^2 + z^2).$$

Thus $y^2 + z^2$ must be a square, say $(y + r)^2$. Then $y = (z^2 - r^2)/2r$. Thus

$$(x, y, z, t) = \left(\frac{z^3 - r^2z}{z^2 + r^2}, \frac{z^2 - r^2}{2r}, z, \frac{z^8 + 14z^4r^4 + r^8}{4r^2(z^2 + r^2)^2} \right).$$

This example has been chosen for three reasons. First, it is of great historical interest. To this problem Fermat appended a note: “Why does Diophantus not ask for the sum of *two* biquadrates to be a square? This is, indeed, impossible. . . .” Later, Euler conjectured that it was also impossible to find three fourth powers whose sum was a fourth power; i.e., to replace t^2 by t^4 . This question remains unsolved.

Second, the problem indicates what happens when the notation is insufficient. First, the chosen unknown is x ; m, y, z are assigned specific values to indicate that they play the role of parameters. But the problem cannot be completed, so the author turns to a sub-problem in which y is the unknown.

Finally, the problem contains a curious case of indifference to sign. The quantity $x^2 - m$ is, in fact, the negative root of $t^2 = (x^2 - m)^2$. Since only the square is used, no harm is done but we must remember that, to Diophantus, the quantity $x^2 - m$, which he used, did not exist. The reader may find it interesting to see why $x^2 + m$ was not used by trying it. Why $m - x^2$, which is positive and produces the same result, was not preferred is a matter for conjecture.

VI. 19: Find a right triangle such that its area added to one of its legs is a square while the perimeter is a cube.

First form the triangle

$$(2x + 1, 2x^2 + 2x, 2x^2 + 2x + 1).$$

The perimeter is $4x^2 + 6x + 2 = (4x + 2)(x + 1)$. Since it is difficult to make a quadratic a cube, consider in turn the triangle

$$((2x + 1)/(x + 1), 2x, 2x + 1/(x + 1)),$$

obtained by dividing through by $x + 1$. The perimeter is $4x + 2$ and the area is $(2x^2 + x)/(x + 1)$. Adding $(2x + 1)/(x + 1)$ to the latter we have $2x + 1$. Thus $4x + 2$ is required to be a cube and $2x + 1$ a square. The obvious value for $2x + 1$ is 4. Thus $x = 3/2$ and the triangle is $(8/5, 3, 17/5)$.

It is not clear whether or not Diophantus implies the more general solution

$$2x + 1 = 4r^6, \quad x = (4r^6 - 1)/2;$$

probably not.

This problem is illustrative of the rather peculiar problems considered throughout *Book VI* and of the complete freedom from geometrical considerations. To Euclid such phrases as “the sum of one side and the area” would have been shocking nonsense.

6 An approximation problem

In *V. 9* it is required as a sub-problem to find two squares, both exceeding 6, whose sum is 13. Since we have, in the first example of the preceding section, a general method of partitioning a number into two squares when one such partition is given, it is merely necessary to set the two values equal, solve for the parameter and approximate this solution in rationals. If this is done with $a = 2$ and $b = 3$, we find $r = 5 + \sqrt{26}$. Approximating by $r = 10$,

$$13 = (258/101)^2 + (257/101)^2$$

and it is readily seen that the conditions are met.

Of course, Diophantus could not do this since the parameters were not expressed. He first finds a number slightly greater than $\sqrt{13/2}$. $13/2 = 26/4$; if $\sqrt{26} \leq 5 + 1/x$, $x^2 < 10x + 1$; let $x = 10$, then $\sqrt{13/2} \sim 51/20$. Now

$$51/20 = 3 - 9/20 = 2 + 11/20.$$

Thus we wish to find a number near $1/20$ such that

$$(3 - 9y)^2 + (2 + 11y)^2 = 13.$$

Then $y = 5/101$ and the squares are precisely those obtained above.

The problem is typical of the approximate methods used. To approximate the n th root of a rational, first write it in the form p/q^n by multiplying numerator and denominator by the necessary integer to make the denominator a perfect n th power. Then multiply p by the n th powers of successive integers until pa^n is sufficiently close to a perfect n th power, say b^n . The approximation is then b/qa . To improve an approximation a_1 to $\sqrt[n]{a}$, set $(a_1 + 1/x)^2 = a$ and approximate x .

7 Transmission of Diophantus

When the Arabs overran the Southeastern Mediterranean in the 7th century, they came into possession of manuscripts of works which had been published

in sufficiently large editions to survive the wars attendant on the breakup of the Roman Empire and the lack of interest in learning of the early Christians. Among these was the *Arithmetic* or at least a portion of it. Translations and commentaries were published in Arabic. These have all been lost; their only trace is in bibliographers' references. When the Arabs formulated their own algebra, they apparently appealed directly to the basic Oriental tradition previously cited. The beginnings of an algebraic notation and the abstract numbers are nowhere to be seen. With the sole exception of the problems mentioned in Section 3 as common to the whole ancient world (Diophantus *I*, 27–30) not one problem from the *Arithmetic* is found in the algebra of Al-Khwarizmi or, as far as is known, in any other basic Oriental text [13]. Probably the Arabs found Diophantus too impractical for their utilitarian mathematics and the Hindus, if they ever saw the *Arithmetic*, were interested in other problems such as the theory of linear indeterminate problems.

In the other reservoir of learning, Byzantium, the manuscripts of Diophantus lay almost unnoticed for eight centuries. We do not know when the missing books were lost but the part which we now possess escaped the sack of Constantinople by the Crusaders in 1204 and later in the same century M. Planudes and G. Pachymeres wrote commentaries on the first part of the *Arithmetic*. At some time, probably in the course of the emigration of the Byzantine scholars during the Turkish conquests, copies were brought to Italy and Regiomontanus saw one there between 1461 and 1464.

The first translation to Latin was made by W. Holzmann, who wrote under the Greek version of his name, Xylander. This translation was published in 1575. Meanwhile, Bombelli, in 1572, distributed all the problems in the first four books among problems of his own in a text on algebra. Bachet, borrowing liberally from Bombelli and Holzmann, made another translation in 1621 and a second edition was published in 1670 including Fermat's marginal notes. In the next two centuries various translations were made into modern languages which were based primarily on the editions just mentioned by Holzmann and Bachet. Finally, in 1890, P. Tannery prepared a definitive edition of the Greek text with a translation into simple mathematical Latin using modern numerical and algebraic notation. From this work the three excellent translations listed in the bibliography have been prepared. The references to the last two paragraphs are the commentaries in [2] and

[6], particularly the latter which has been followed rather closely.

Bibliography

A. Greek Text with Latin Translation:

1. C. G. Bachet, *Diophanti Alexandrini Arithmeticonum libri sex, etc.* Paris, second edition, 1670.
2. P. Tannery, *Diophanti Alexandrini opera omnia cum Graecis commentariis*, Teubner, vol. i, 1893, vol. ii, 1895.

B. Modern translations based on [2]:

3. A. Czwalina, *Arithmetik des Diophantos aus Alexandria*, Göttingen, Vandenhoeck, 1952.
4. P. Ver Eeke, *Diophante d'Alexandrie. Les six livres arithmétiques et le livre des nombres polygones*. Bruges, Descée, 1926.
5. T. L. Heath, *Diophantus of Alexandria*, Cambridge, 1910.

C. Histories and Compilations of Ancient Mathematics:

6. T. L. Heath, *History of Greek Mathematics* (two volumes), Oxford, 1921.
7. —, *A Manual of Greek Mathematics*, Oxford, 1931.
8. G. Loria, *Le Scienze esatte nell' antica Grecia*, 2nd edition, Milan, Hoepli, 1914.
9. G. H. F. Nesselmann, *Die Algebra der Griechen*, Reimer, Berlin, 1842.
10. O. Neugebauer, *The Exact Sciences in Antiquity*, Princeton, 1952.
11. I. Thomas, *Selections Illustrating the History of Greek Mathematics*, Harvard and Cambridge (Loeb Classics) 1939.
12. B. L. van der Waerden, *Science Awakening*, P. Noordhoff, Groningen, 1954 (English translation by A. Dresden).

D. Miscellaneous References:

13. S. Gandz, "The Origin and Development of the Quadratic Equations in Babylonian, Greek and Early Arabic Algebra", *Osiris*, vol. 3 (1938) pp. 405–557.
14. J. Klein, "Die griechische Logistik und die Entstehung der Algebra", *Quellen und Studien zur Ges. der Math. Abt. B*, vol. 3, 1934–6, pp. 18–105; 122–235.
15. O. Neugebauer, *Mathematische Keilschrift-Texte*, 3 Vols., Springer, 1935–7 = *Quellen und Studien zur Ges. der Math. Abt. A*, vol. 2.
16. O. Neugebauer and A. Sachs, *Mathematical Cuneiform Texts*, American Oriental Society, New Haven, 1945 = *Am. Oriental Series*, vol. 29.
17. *Paulys Real Encyclopädie der Classischen Altertums Wissenschaft*, Stuttgart, 1893.

Hypatia of Alexandria

A. W. RICHESON

National Mathematics Magazine 15 (1940), 74–82

The first woman mathematician regarding whom we have positive knowledge is the celebrated mathematician-philosopher Hypatia. The exact date of her birth is not known, but recent studies indicate that she was born about A.D. 370 in Alexandria. This would make her about 45 years of age at her death. Hypatia, it seems, was known by two different names, or at least by two different spellings of the same name; the one, Hypatia; the other, Hyptachia. According to Meyer [6], there were two women with the same name living at about this time; Hypatia, the daughter of Theon of Alexandria; the other, the daughter of Erythrios. Hypatia's father was the well-known mathematician and astronomer Theon, a contemporary of Pappus, who lived at Alexandria during the reign of Emperor Theodosius I. Theon, the director of the Museum or University at Alexandria, is usually considered as a philosopher by his biographers.

Hypatia's biographers have given us but little of her early personal history. We know that she was reared in close touch with the Museum in Alexandria, and we are probably safe in assuming that she received the greater part of her early education from her father. If we are to judge from the records which the historians have left us, we would conclude that her early life was uneventful. It would seem that she spent the greater part of her time in study and reading with her father in the Museum.

Suidas [9] and Socrates [8], as well as others who lived at the same time, lead us to believe that Hypatia possessed a body of rare beauty and grace. They attest not only to her beauty of form and coloring, but each and every one speaks just as highly of the beauty of her character. In the absence of a life painting of Hypatia we must depend upon the conception of others for a picture of the philosopher. In the intro-

duction to his edition of Theon's *Commentary* [3] Halma has given us a short biography of Hypatia. On the title-page there is a medallion which gives his conception of the philosopher [see the next article]. Meyer feels that this drawing is unfortunate, as he does not believe it gives a true impression of the woman Hypatia. Charles Kingsley, on the other hand, in his novel *Hypatia*, has written a vivid description of his impression of the philosopher.

If we are to believe the historians as to her beauty, we would expect that she was eagerly sought after in marriage. This apparently was the case: her suitors included not only outsiders, but many of her students as well. The question of her marriage, however, leads us to one of the controversial points of her life. Suidas states she was the wife of the philosopher Isidorus; then 25 lines later, he states she died a virgin. This apparent contradiction has been explained in several ways by later writers.

Toland [11] believes she was engaged to Isidorus before she was murdered, but was never married. Hoche [4] is of the opinion that the mistake arose from Suidas' abstract of the works of Damascius, a conclusion which Meyer does not believe to be true, pointing out that he found on the margin of one of Photius' works the statement, "Hypatia, Isidore uxor." Since Photius transcribed Hesychius' works, it is possible that the error arose in this manner. The evidence against such a marriage is further substantiated by the fact that Damascius states that Isidorus was married to a woman named Danna and had a child by this wife. Another fact which should be taken into consideration is that Proclus was much older than Isidorus: it has been pretty definitely established that Proclus was born about 412, and, since Hypatia's death occurred in the year 415, it would be impossible for Hypatia to have been the wife of

Isidorus. The present writer is inclined to agree with Meyer that the mistake arose in Photius' transcription of Hesychius' work and that Hypatia was not married at any time in her life.

The second controversial point is the question of her death. In studying the statements made by many of the historians in regard to her death it seems desirable to review the murder in relation to the events which had happened previously. It is necessary for us to investigate not only Hypatia's relation to paganism, but also the relation between Cyril, the Christian bishop at Alexandria at this time, Orestes, the Roman Governor at Alexandria, and Hypatia. In view of the triangular relationship, we shall recall briefly some of the important events just prior to and during the episcopate of Cyril and their relationship to the authority of the Roman Governor.

On October 12, 412, Theophilus, the Bishop at Alexandria, died, and six days later his nephew Cyril was elevated to the episcopate of Alexandria. From the outset the new bishop began to enforce with zeal the edicts of Theodosius I, the Roman Emperor, against the pagans, along with restrictions which he himself promulgated against the Jews and unorthodox Christians. He further began to encroach upon the jurisdiction which belonged to the civil authorities; that is, to the Roman Governor. It must be remembered that the population of the city of Alexandria in the fourth and fifth centuries of the Christian era consisted of a conglomeration of nationalities, creeds, and opinions, and that nowhere in the Empire did the Romans find a city so difficult to rule as Alexandria. The people were quick-witted and quick-tempered, and we read of numerous clashes, street fights, and tumults, not only between the citizenry and the soldiers, but also between the different classes of citizens themselves. There were frequent riots between the Jews and the Christians on the one hand and the pagans and the Christians on the other. The Christian population did little or nothing to quiet these people, but even added one more controversial topic for them to quarrel about. Consequently we find that the edicts and promulgations of Cyril not only caused strife among the people but aroused the anger of the Roman Governor, Orestes, the one person who stood in the way of the complete usurpation of the civil authority by Bishop Cyril. Friction continued between these two until there was a definite break in their relations.

Because of her intimacy with Orestes, many of the Christians charged that Hypatia was to blame, at least in part, for the lack of a reconciliation be-

tween Orestes and Cyril. Socrates states that some of them, whose ringleader was named Peter, a reader, driven on by a fierce and bigoted zeal, entered into a conspiracy against her. They followed her as she was returning home, dragged her from her carriage, and carried her to the church Caesareum, where they stripped her and then murdered her with shells. They tore her body to pieces, took the mangled limbs to a place called Cinaron, and burned them with rice straws. This brutal murder happened, he says, under the tenth consulate of Honorius and the sixth of Theodosius in the month of March during Lent, so that the year of her death may be set as 415.

Socrates' report of Hypatia's death is corroborated not only by Suidas, but also by other historians such as Callistus [1], the ecclesiastical historian, Philistorgus [12], Hesychius [2] the Illustrious, and Malalus [5]. Damascius says that Cyril had vowed Hypatia's destruction, while Hesychius states that his envy was caused by her extraordinary wisdom and skill in astronomy. Damascius also relates that at one time Cyril, passing by the house of Hypatia, saw a great multitude, both men and women, some coming, some going, while others stayed. When he was told that this was Hypatia's house and the purpose of the crowd of persons was to pay their respects to her, he vowed her destruction.

When we compare these statements, it would seem that Hypatia's death, or at least the occasion of it, was due to her friendship with Orestes. This friendship enraged the Christian populace because they felt that she prevented a reconciliation between Cyril and Orestes. We are also led to believe that the more sober-minded of the Christians yearned for a reconciliation between these two and that no doubt her death was ordered by Cyril.

Among the later writers on the subject there is a divergence of opinion. Toland lays the death of Hypatia directly at the feet of Cyril. Wolf [13], on the other hand, is inclined to believe that Cyril knew beforehand that the murder was being plotted but did nothing to prevent it. As to the causes of the murder, Wolf mentions her belief in paganism and her teaching of Neoplatonism, along with the practice of treating the mentally diseased with music, all of which might be considered as coming under the pale of the edicts of Theodosius I regarding pagan worship.

The present writer is inclined to follow Meyer part of the way in the interpretation of these events; that is, Hypatia was used as a sacrifice for a political or personal vengeance, possibly a political vengeance.

Cyril and Orestes were at odds; both had made various reports to the Emperor, each one attempting to show that his actions were justified. On the other hand, Orestes was the one person who stood in the way of the complete assumption of the civil power by Cyril, and naturally Cyril was eager to use every incident which would embarrass Orestes. In the case of Hypatia's death it would seem that its underlying cause was not so much a struggle for the assumption of the civil authority, but rather a struggle of the Christian church against the pagan society of Alexandria. It must be remembered that although Orestes professed Christianity, the fact still remained that this profession was more one of policy than of faith. In all justice it would certainly seem that Cyril should be held at least indirectly responsible for her death. Certainly he could have prevented the mob's violence, if he had made the slightest effort.

Meyer feels the relation between Cyril and Synesius should be considered in investigating Hypatia's death. He is of the opinion that possibly there was an old difference between these two, and that her death was brought about by Cyril in order to settle this difference with Synesius. Meyer bases his conclusions on the contents of Epistle 12 [10] of Synesius, in which he exhorts Cyril to go back to the Mother Church, from which he had been separated for a period of time for the expiation of sin. The present writer is of the opinion that Meyer has no justification for this assumption. Although we do not know the exact date of Synesius' death, it was probably between 412 and 414, and it must be remembered Cyril was not raised to the bishopric until late in the year 412. It is very probable that Epistle 12 was written before Cyril was made Bishop at Alexandria, though as a matter of fact we have no convincing evidence that the letter was written to Saint Cyril. Furthermore, there is no evidence to support the belief there ever existed any difference between Cyril and Synesius.

It has been stated above that little is known concerning Hypatia's early life. Consequently there is little on which to base our conclusions regarding her early education. It goes without saying that her father taught her in mathematics, astronomy, and science. Beyond this we do not know who her teachers were, but we rest assured that, with an intellect as fertile as hers, she was not long satisfied with the narrow training in mathematics and astronomy. In order to understand the possible trend of her education it is necessary to take a look at the working of the Museum at Alexandria. The Museum had its origin in

the efforts of Ptolemy Soter about 300 B.C., when he brought to the city of Alexandria all the philosophers and writers it was possible for him to obtain. To these he gave every encouragement possible, not only financial aid, but also in books and manuscripts from Greece. The later rulers of Egypt continued their support until the country came under Roman authority in 30 B.C. This ended the first period of intellectual activity, which is characterized as purely literary and scientific in nature. With the conquest of the country by the Romans, intellectual activity was again in the ascendancy and Roman, Greek, and Jewish scholars were again attracted to the city. This second school of thought was somewhat different from the first. We have an intermingling of nationalities with their varying philosophies and personalities, all of which developed into the speculative philosophy of the Neoplatonists, the religious philosophy of the early Christian fathers, and the gnosticism of the Oriental philosophers. This second period of intellectual activity continued until about 642, when the city was destroyed by the Arabs. Considered as a whole, the Alexandrian School stood for learning and cosmopolitanism, for erudition rather than originality, and for a marked interest in all literary and scientific techniques. It was at the Museum that these philosophers, writers, and scientists gathered to lecture to their students and to converse with one another. Theon, Hypatia's father, was director or fellow in the Museum, and it is reasonable to infer that Hypatia came into close contact with the leading educators and philosophers of Alexandria.

The question is frequently asked whether or not Hypatia studied at Athens. Here again we come to a point which has not been definitely decided. Suidas says she obtained part of her education there, or at least the passage has been so interpreted, for both Meyer and Hoche are of the opinion that Suidas has been misinterpreted on this point. Wolf states that Hypatia studied at Athens under Plutarch but Meyer again points out that this was highly improbable, as at the time Plutarch was lecturing at Athens, Hypatia was probably 30 years of age and was herself lecturing at Alexandria. Suidas also makes mention of the fact that she studied under another philosopher at Alexandria, but he does not identify this philosopher except to say that it was not Theon. Meyer thinks it might have been Plotinus. Regardless of how or where she received her education, we do know that she received a thorough training in arts, literature, science, and philosophy under the most competent teachers of the time.

It was with this training that she succeeded to the leadership of the Neoplatonic School at Alexandria. The exact date at which she assumed control of the school is not known, but Suidas informs us that she flourished under Arcadius, who was Emperor of the Eastern Roman Empire from 395 to 408. We are naturally led to ask two questions regarding her teaching: first, what was her ability as a teacher? second, what was the nature of her teaching? The first question is much simpler than the second, although there are sufficient facts relating to the nature of her teaching to enable us to draw a fairly definite conclusion.

All the contemporary and later writers on this period testify to the high reputation of her work as a teacher. Each one attributes an extraordinary eloquence and an agreeable discourse to her lectures. Suidas speaks highly of her teaching methods, while Synesius in one letter praises her voice and in another mentions that her philosophy was carried to other lands. Socrates and Philistorgius tell us that not only the Egyptians, but students from other quarters of Europe, Asia, and Africa came to her classes until there was in reality a friendly traffic in intellectual subjects. Suidas states that, on account of her ability as a teacher and her personality, Orestes sought out her house to be trained in the art of public manners. Damascius states she far surpassed Isidorus as a philosopher, and it should be remembered that Damascius was a friend and pupil of Isidorus.

Among her disciples there are many well-known men other than Synesius. The names of these include Troillius, the teacher of the ecclesiastical historian Socrates, Euoptius, the brother of Synesius and probably the Bishop of Tolemais after the death of Synesius, Herculianus, Olympius, Hesychius, and finally Herocles, the successor of Hypatia in the Platonic School at Alexandria.

From her teaching position she expounded the philosophy of the Neoplatonic School and her fame rests primarily upon the manner in which she conducted this school. In her teaching she no doubt lectured not only on philosophy as we know it today, but also included the scientific subjects of mathematics, astronomy, and the subject of physics as known at the time. She was apparently well versed in astronomy, since Suidas tells us that she excelled her father in this field. We may also assume that she taught the rudiments of mechanics, since there is a reference in one of Synesius' letters to an astrolabe which she constructed, and in another letter Synesius requests Hypatia to make a hydroscope for him.

Neoplatonism, as a philosophic system of thought, had its inception during the second century of the Christian era. It was built up from the remains of many of the systems of philosophies of ancient Greece and became a religion for many of the heathens, who could no longer believe in the old gods of Olympus. The Neoplatonist believed in a supreme being or power, which was the Absolute or One of the system. This supreme power was mystic, remote, and unapproachable in a direct fashion by finite beings. Hence there existed between man and the Absolute lesser gods or agencies. The first in this series was Nous or Thought, which was emanated by the Absolute as an image of itself. Below Nous, there existed the triad of Souls, which pervaded all of the material universe, and all of those beings with which it is peopled are a direct emanation from the triad of Souls. Matter or material things were thought of as belonging to an evil category, while the triad of Souls belonged to a pure category. Man, a mixture of the material and the spiritual, has the power by indulging in self-discipline and subjugation of the senses, to lift himself to a level where he may receive from the Absolute a revelation of divine realities. Once man has caught a glimpse of this vision, he is able to free himself entirely from the thralldom of matter.

It should be noted that the development was from a higher to a lower or descending series. Since each series participated in the one above it, there was also a turning back, where the soul by an ascending process was able to return to the Absolute. The object of life, when the soul was perfectly free, was to rise by the practice of virtue from the category of matter to the higher category of intelligible realities. There were purifying virtues, which disciplined the soul till it became capable of union with the Absolute.

We have no writings of Hypatia, but we may rest assured that she at least subscribed to the general principles of Neoplatonism. Plotinus' works show that he succeeded in contempt of bodily cares and needs, and we find the same thing to be true with Hypatia. No doubt Hypatia's use of logic, mathematics, and the exact sciences gave her a discipline which kept her and her pupils from going too far in the superstitions and speculations of some members of this group of thinkers. Synesius in his speech before the Arcadians, acknowledges the purely subjective character of the different attributes which are conceived of by man as belonging to the divine nature. He also felt a wholesome reticence in his attempts to reach towards the Incomprehensible. He believed in the

Trinity of Plotinus, but did not assign to the World-soul the creating or animating of the entire universe. He thought occasional supernatural communications between God and the human soul were possible, and he also believed that man was able to purify his soul to such an extent that he would be able to elevate the imagination to a point where it would be possible for him to share in the ecstasy of the upper light. He believed that the final goal aimed at in life was a pure and tranquil state of mind, undistracted by fierce passions, gross appetites, or the demands of worldly affairs. It would be reasonable to assume that these tenets of Synesius' faith were inculcated in him by his beloved teacher Hypatia.

In considering the writings of Hypatia we have but little information to fall back on. Suidas is the only historian to give us any information concerning her writings. He gives us the names of three: a commentary on the *Arithmetica* of Diophantus of Alexandria, a commentary on the *Conics* of Apollonius of Pergassus, and a commentary on the *Astronomical Canon* of Ptolemy. None of these are extant at the present time.

We are naturally led to the question of why Hypatia, a student of philosophy, a teacher of renown, and the leader of the Neoplatonic School at Alexandria, left only three works and those three purely mathematical or astronomical. The answer is probably that Suidas quoted the writings of Hypatia as given by Hesychius, who for some reason gives an account only of the Astro-Mathematical works of Hypatia. It is rather difficult for us to believe that with approximately twenty years of teaching she would produce not more than three works, and those three commentaries. So we are led to the conclusion that Hypatia did leave other writings, which were probably lost in the destruction of the library at Alexandria, and that these works were principally philosophic in nature. It is true that both Halma and Montucla [7] make mention of other works of Hypatia; Halma in particular says she left behind "beaucoup d'écrits." At the present time it is impossible to determine from what source Halma obtained this information, and it is more than probable this is only a conjecture on his part.

With the passing of Hypatia we have no other woman mathematician of importance until late in the Middle Ages. Although we have no definite information to indicate that she exerted any great influence on the development of mathematics or science in general, nevertheless she certainly passed on to her scholars and followers a discipline and restraint which were carried over to a later period. It is possible that the effects of her teachings have been lost sight of, since any works she might have left behind were certainly lost when the Arabs destroyed the Library at Alexandria in 640.

References

1. Callistus, *Nicephori Kallisti historia ecclesiastica Migne, Patrologiae Graecae*, Tome 147, Paris, 1856.
2. J. Flacch, ed., *Hesychii Milesii Onomatologie que supersunt cum prolegomenis*, Leipzig, 1882
3. Halma, ed. *Theon d'Alexandrie, Commentaire sur le livre III de l'Almageste de Ptolemée*, Paris, 1882.
4. Richard Hoche, Hypatia die Tochter Theons, *Philologus*, Fünftehnter Jahrgang, Göttingen, 1869, pp. 435–474.
5. Johannus Malalus, *Chronographia ex recensione Ludovici Dindorfii*, Bonn, 1831.
6. Wolfgang Alexander Meyer, *Hypatia von Alexandria*, Heidelberg, 1886, p. 52.
7. J. F. Montucla, *Histoire des Mathématiques*, Tome I, Part I, Liv. V, Paris, 1799.
8. Socrates, *The Ecclesiastical History*, trans. by Henry Bohn, London, 1853.
9. Suidas, *Lexicon, Lexicographi Graeci*, Vol. I, Pars. IV, ed. by Ada Adler, Leipzig, 1935.
10. Synesius, *Opera quae extant omnia, Patrologiae Graecae*, Tome 66, Paris, 1864.
11. John Toland, *Tetradymus*, London, 1720, pp. 101–136.
12. H. Valesio, ed., *Ex ecclesiastici Philostorgii historia epitome confecta a Photia patriarcha*, Paris, 1873.
13. Stephen Wolf, *Hypatia, die Philosophin von Alexandrien*, Vienna, 1879.

Hypatia and Her Mathematics

MICHAEL A. B. DEAKIN

American Mathematical Monthly 101 (1994), 234–243

1 Introduction

The first woman mathematician of whom we have reasonably secure and detailed knowledge is Hypatia of Alexandria. Although there is a considerable amount of material available about her, very much of that is fanciful, tendentious, unreferenced or plain wrong. These limitations are to be found even in works that we might hope to be authoritative; for example, the entry in the *Dictionary of Scientific Biography* (DSB) [11]. Even where the account given is more careful and accurate [14, 19, 20], one is disappointed to be told so little of Hypatia's *Mathematics*.

This article will direct the reader's attention to the best accessible sources and will describe what is known about her mathematical activities.

2 The historical background

In about 330 B.C., Alexander the Great conquered northern Egypt and, via a deputy (Ptolemy I Soter), founded a city (Alexandria) in the Nile delta. This almost immediately became home to the Alexandrian Museum, an institution of higher learning, rather akin to the medieval universities of some 1500 years later. Euclid was an early (probably the first) "professor" of mathematics.

The Museum continued for many centuries. In 30 B.C., Cleopatra's suicide allowed the Roman Empire to occupy Alexandria, but this event destroyed neither the city's Greek heritage nor its intellectual tradition. In the years that followed, two of the greatest of the late Greek mathematicians flourished in Alexandria. Diophantus was active around A.D. 250 and produced in particular his *Arithmetica* at this time. Several generations later, Pappus (c. 300–c. 350) also worked there.

A later mathematician, Theon of Alexandria, was the last person definitely known to have been associated with the Museum. Because he recorded two eclipses (one of the sun and one of the moon) and because he is also credited with achievements during the reign of Theodosius I, it is thought that he was at the height of his powers in the decade 360–370. Theon may well have been the last "president" of the Museum. His daughter, Hypatia, was associated with the Neo-platonic School — a different institution.

Alexandria, in the years around A.D. 400, was a turbulent mix of cultures. Christians were in the majority, but they were divided among themselves.



Gasparo's portrait of Hypatia

There were also persons whom the Christians regarded as “pagans”; these could be anything from believers in the Olympian pantheon to adherents of various schools of “Neoplatonic” thought. Beyond these there were also Jews and Gnostics.

The Roman Empire, of which Alexandria was a part, was under external pressure from the Huns and the Visigoths. It split in 395 into the Western Empire (ruled from Rome) and the Eastern Empire (ruled from Constantinople). The official religion was Christianity: it had been established under Constantine. But there had been relapses; in particular, Julian the Apostate had reigned over the combined empire from 361 to 363.

At the time of Hypatia’s death, the local governor was Orestes, a Christian not unsympathetic to other views, but whose authority was under challenge from that of the less tolerant Cyril (St. Cyril of Alexandria) who acceded to the bishopric in 412. The divisions that beset the city were prone to erupt into sectarian violence; the great libraries associated with the Museum were one by one destroyed, the last going up in smoke in 392 when the temple of Serapis was put to the torch during a riot. Another such disturbance was to claim Hypatia’s life in the second decade of the fifth century. She died, brutally hacked to pieces, at the hands of a Christian lynch-mob.

Following this, very possibly in part because of it, the thrust of Neoplatonist thought and education moved from Alexandria to Athens. Three names require mention. Proclus (410?–485) was the last of the great mathematicians of antiquity. He frequented the Neoplatonic School at Athens and is best remembered for a commentary on Book I of Euclid’s *Elements*. After Proclus came Isidorus and his pupil Damascius (philosophers both of them rather than mathematicians, although the latter may have some claim on a place in mathematical history [6, pp. 312–313]). In 529, the emperor Justinian, enforcing Christianity as the state religion, closed the Neoplatonic School and Damascius went into exile in Persia.

3 The primary sources

The oldest accounts of Hypatia come to us from either the *Suda* (or *Suidae*) *Lexicon* or from the writings of the early Christian Church. For an accessible account of them, giving more detail than I provide here, see Mueller [14].



Medallion of Hypatia in the Introduction to Halma’s edition of Theon’s *Commentary on the Almageste*. (Artist Unknown)

Briefly, the *Suda* was a 10th-century encyclopedia, alphabetically arranged, and drawing on earlier sources. In the case of Hypatia, these are in part known. (One is a now lost work, a life of Isidorus by Damascius.) The relevant entry is unusually long, but is not seen as reliable in all its aspects (see [25]); indeed in places it contradicts itself.

The other sources are to be found in the main in a compilation known as the *Patrologiae Graecae* [13], or PG for short. This gives earlier accounts (particularly of her death) than are available in the *Suda* and also preserves letters to her and about her from the hand of one of her pupils, Synesius of Cyrene. Also by Synesius is a letter published as a separate document included with the others in FitzGerald’s translation [4].

4 Life and legend

The best-recorded event in Hypatia’s life is her death and the manner of it. The fullest account tells us that a crowd of Christian zealots led by one Peter the Reader seized her, stripped her and proceeded to dismember her and burn the pieces of her corpse. Another says she was burned alive, but this would seem to be a less accurate version.

The political background to this action has been the cause of much speculation. Gibbon [5] is by no means alone in attributing the guilt for the murder to Cyril, but Rist [20] disputes this, which does mean taking issue with the *Suda*. Rist’s account, in essence, has it that, like victims of violence in Belfast or Beirut today, she was seized not with any great selectivity at all, but rather because she was a well-known public figure, prominent on the other side of a religious divide. This to my mind is quite compatible with the statement quoted by Gibbon to the effect that she was killed because of her outstand-

ing ability. We need not posit any specific jealousy to say this, and Rist thinks it is unlikely that precise differences of doctrine led to her death. Rist does toy with the idea that her mathematical activities were a partial cause, hypothesizing that these included astrology. This, to me, sets us on a path we have no reason to travel.

The *date* of her death is now generally accepted to have been 415, although others have been suggested. See Mueller [14] for details.

The date of her birth is much less certain. (This is to be expected — people are not, generally speaking, famous when they are born.) The eclipses described by Theon, Hypatia's father, have been dated to 364. So, from the eclipses to the time of her death is an interval of 51 years. Valesius, an early commentator on the PG who had the wrong date for the eclipses, reckoned this interval at 47 years; rounding this to 45 produces a date of c. 370, which is the generally-stated figure. Of course, astrology aside, we have no real reason to suppose that her birth coincided with the eclipses; nor have we any idea how old Theon (or more importantly his wife) was in 364. (I tend to agree with Mueller that a date of c. 350 is more plausible.)

As to her life between these uncertain dates, we may readily summarize. She was a respected and eminent teacher, charismatic even, and beloved of her pupils (e.g., Synesius). We have evidence that she was regarded as physically beautiful, that she wore distinctive academic garb, that she taught not only Mathematics but also Philosophy, that she gave public lectures and may have held some kind of public office.

She seems to have been determinedly celibate, indeed repelling one ardent suitor by confronting him with one of her used menstrual pads and lecturing him on the shameful and unclean nature of what he thought beautiful (the vagina).

Although almost all the primary sources are Christian and tell of the life and death (at Christian hands) of a prominent advocate of a rival philosophy, they do so in such a way that we are left with a favorable impression of her. My reading of this is that the official discouragement of her teachings on the part of the Church authorities and of their (Christian) civic counterparts was far from complete.

Certainly that favorable impression has informed various works of literature of which the best-known in English are Kingsley's novel [10] and the passage from Gibbon. Also fiction is Hubbard's telling of Hypatia's story [9]. It formed a chapter in a popular

reader early this century and has given us the most widely disseminated "portrait" of Hypatia, attributed to an artist called Gasparo, of whom I am able to learn nothing. (Of course such "portraits" have exactly the same validity as (e.g.) Doré's illustrations of the Bible.)

5 Hypatia's Philosophy

The Philosophy expounded by Hypatia is known to have been Neoplatonist. There were various versions of Neoplatonism, all endowing Plato's Theory of Forms with an explicitly religious dimension. Richeson [19] describes one such system; Rist [20] suggests that Hypatia actually preached another.

Richeson does however make a particularly insightful remark on the connection between Neoplatonist Philosophy and Mathematics. The nature of Mathematics is to abstract — to derive *ideas* from material things. Thus Geometry, although it has its *origin* in the practical world of land surveyors and inspectors of weights and measures, transcends these beginnings. The *Elements* deals with a world that is no longer the world of the practical but rather the world of ideas. Thus Mathematics could be seen as a paradigm of that transcendence over the material that Neoplatonism enjoined.

6 Hypatia's Mathematics

That Hypatia was a mathematician is beyond doubt. The PG tell us that she learned her Mathematics from her father Theon and went on to excel him in the subject and to teach it to numerous students. Another such source is more critical: "Isidorus greatly outshone Hypatia, not just because he was a man and she a woman, but in the way a genuine philosopher will over a mere geometer." This opinion, which will earn no praise from either women or mathematicians, is thought to derive from Damascius' life of Isidorus, the lost work that in part informed the *Suda*. (Marrou [12], following Tannery [25], supplies the following delightful gloss: "[it] means in plain language that Isidorus knew nothing of mathematics.")

However, the *Suda* itself gives the most explicit account of Hypatia's mathematical works. It attributes to her the authorship of three works. The only things she is known to have written all deal with Mathematics or Astronomy. The books that many feel she must have authored on Neoplatonist Philosophy receive no mention. Others (e.g., Kramer [11])



have credited her with further works of Mathematics. For this there is no evidence, except in one specific instance to be described below. The relevant passage in the *Suda* is precisely twelve words long. And even this short excerpt is the subject of various alternative and disputed readings. However, there is a general consensus that Tannery [25] is correct in rendering it thus: “She wrote a Commentary on Diophantus, [one on] the astronomical Canon, and a Commentary on Apollonius’s *Conics*.”

“Commentaries” were what we would now refer to as “Editions” (with the obvious difference that they needed to be copied by hand), and the author of a “Commentary” is perhaps better referred to as an “Editor.” Such “Editors” or “Commentators” did, however (to a greater or lesser extent, and with greater or lesser care to distinguish their own contributions from the original), provide new material of various sorts (witness Fermat’s famous marginal note to Diophantus). It should be noted that in many cases the original text has come down to us only through Commentaries or translations (often into Arabic).

Theon, Hypatia’s father, was a prolific author of Commentaries. He wrote one on the *Elements* (which, in places, still provides our present text), on two other works by Euclid, the *Data* and the *Optics*, and on two works by Ptolemy, the *Almagest* and the *Handy Tables*. There were also works now lost or partly so; particularly germane to our story is a work on the astrolabe. For this and more, see Toomer [28].

The picture that emerges of Theon is one of an editor, teacher and textbook-writer rather than a research mathematician. So is he judged, often with more than a hint of disapproval. But this should not mean that his was a wasted life. His works were preserved, presumably because they were perceived as having lasting value. It is all too understandable, given the politics of late 4th-century Alexandria and the decay of the Museum, that the emphasis on research (possible in Pappus’s time) should be replaced by the priority of conserving knowledge.

After considering her works *seriatim*, I shall offer the hypothesis that in her scholarly priorities Hypatia was very much her father’s daughter. This, as I hope I have just made clear, is not to denigrate her.

7 Apollonius’ *Conics*

Apollonius lived around 200 B.C. and the *Conics* is the most important of his surviving works. See, for more detail, Toomer’s account [27]. There are very few sources for our present text and Hypatia’s *Commentary* is not one of them. Of the eight books that make up the *Conics*, the first four survive via a *Commentary* by Eutocius while three of the remaining four have come down to us via the Arabic. The other is lost, as is also, we must conclude, Hypatia’s *Commentary*, unless it is the lost original of Eutocius’ work.

8 The Astronomical Canon

In the case of the “Astronomical Canon”, we are much better placed. It is now generally assumed that Tannery’s interpolation (the words in brackets in Section 6) in the *Suda* entry is correct. This means that this work also was a commentary. The most likely original is one of the works of Ptolemy, either the *Almagest* or the *Handy Tables*. It will be remembered that Theon wrote commentaries on both these works.

Theon’s commentary on the *Almagest* has been printed in various editions. The best and most recent is by A. Rome [21, 22]. (But see also [23].) It comprises separate commentaries on the thirteen books that go to make up the *Almagest*. The titular inscriptions (as described by Rome from his study of the manuscripts) of the first and second books ascribe these works to Theon himself. Books 4–13 contain no inscriptions. Only the very best manuscripts contain the *Commentary* on Book Three, and here the

inscription tells us that the work is Theon's "in the recension of my philosopher-daughter Hypatia."

Heath [8], reviewing Rome's work, thus ascribed this chapter of the commentary to Hypatia, with the inference that it was also the work alluded to in the *Suda*, and that Theon (recognizing his daughter's work as superior to his own) had suppressed his earlier effort in favor of hers. (The pity, from our point of view, is that we don't have both versions before us; so we cannot see for ourselves where or how or to what extent Hypatia's commentary differed from Theon's.) Rome himself discusses the matter at considerable length in his later work [22], but in such a way as not to rule out a possibility that has been canvassed: that father and daughter collaborated.

Neugebauer [16, p. 838] accepts this as likely. However, he regards it as probable that what the *Suda* refers to is a commentary not on the *Almagest* at all, but on the *Handy Tables*. This is because the same word (*Canon*) is used for both works. (Delambre [2] had earlier noted this same concordance of wording, but as his work predates Tannery's suggested interpolation, he credits Hypatia with a set of Astronomical Tables.) If the *Suda* were referring to a commentary on the *Almagest*, so the argument goes, then it would speak of the *Syntaxis*, rather than the *Canon*. (*Syntaxis* is the Greek name for the work we now know by its Arabic designation.) Against this, however, is the Canon of Parsimony and the fact that Book 3 of the *Almagest* has a strongly tabular character.

9 Diophantus' *Arithmetic*

We may also have some of Hypatia's own writing from the commentary on Diophantus. Diophantus' major work is the *Arithmetic*, originally comprising thirteen books. Of these only six now survive from the Greek, and possibly part of another, now listed as separate, the *Polygonal Numbers*. Tannery [26] suggested that all existing manuscripts known to him derived from a common source and that that source was Hypatia's commentary. His careful "family tree" of the manuscripts was later modified in one detail and made available in the amended form in Heath's Edition [7]. The presumption was that Books 7–13 are lost because Hypatia's commentary did not include them, much as Eutocius' commentary extended only to the first four books of the *Conics*. This hypothesis enjoyed a deal of support, and Vogel's article on Diophantus in the DSB simply accepts it.

The basis for this theory was the Greek text and the fact that the *Suda* reference to Hypatia's commentary is the only mention of so ancient an edition. Sesiano [24, pp. 71–75] however queries this account. This is a matter of great controversy. The old theory will be presented first, but see the remarks at the end of this section.

On the old story, the mathematical world of today owes Hypatia a great debt, for without her we would have much less of the works of Diophantus. But there is an obvious corollary. If what survives for us is Hypatia's commentary, then some of her work may appear there. It may be possible to see what is hers. One complication is that a later scribe was thought to have attempted to reconstruct Diophantus' original text and thus to have systematically omitted material he judged to be interpolated. But "the distinction between text and scholia being sometimes difficult to draw, he included a good deal which should have been left out" [7, p. 14].

On this account, the most likely of the supposed interpolations to have come from Hypatia's hand are two "student exercises" at the start of Book II. The first asks for the solution of the pair of simultaneous equations:

$$x - y = a, \quad x^2 - y^2 = (x - y) + b,$$

where a, b are known. The next is a minor generalization. It requires the solution of the pair of simultaneous equations:

$$x - y = a, \quad x^2 - y^2 = m(x - y) + b,$$

where a, m and b are known. There is some evidence to link this problem to Hypatia: a nine-word phrase in the original Greek is identical with one from Euclid's *Data*, which her father had edited.

Recent work by Roshdi Rashed, Sesiano and others has suggested that some of the lost books of Diophantus in fact survive in Arabic translations. This has led to very great and indeed bitter controversy. What is at issue (apart from the personal rivalries involved) is whether Diophantus or someone else wrote the newly discovered works and where they might fit into the fragment previously published. Sesiano and others are inclined to the view that if anything of Hypatia's commentary survives then it survives in the Arabic. There are no clear indications of what might be by her and what by Diophantus or by other scholiasts. Many of Sesiano's conclusions are hotly disputed by Rashed [18]. However, tentative attributions of material to Hypatia all tend to

accept the overall assessment reached above — that her contributions to mathematical knowledge itself were slight or non-existent.

10 The astrolabe

The other source for information about Hypatia's mathematical activities is the correspondence of Synesius.

There is a brief but telling reference to Hypatia in Synesius' essay-letter *De Dono Astrolabii*. The name "astrolabe" was a term applied to a variety of instruments. For a good overview of later developments, see [17]; earlier ones are discussed by Neugebauer [15]. A simple attempt to replicate the motions of the heavens in a mechanical model produces the device known as an "armillary sphere." Such an object is necessarily 3-dimensional and unwieldy, more suitable for display purposes than for use as a practical instrument of observation or computation.

However, once we have a theory of stereographic projection, the way is open for the construction of a more practical two-dimensional device. This theory was given by Ptolemy in his *Planisphaerium*, which even includes tabular material. Whether Ptolemy went on to develop the "little astrolabe" (i.e., the practical instrument) has been argued. Neugebauer regards it as probable that he did.

The next figure is Theon. Ptolemy died in about 170 A.D., about two centuries before Theon's active period. Theon wrote, as we have seen, commentaries on the *Almagest* and the *Handy Tables*. The *Suda* also credits him with a treatise on the little astrolabe, and Arab sources refer in addition to a work of his on the armillary sphere. This set corresponds exactly to the set of works assigned to Ptolemy by the Arabs.

There is thus considerable evidence that Theon was familiar with the theory of the little astrolabe. We might speculate that he invented it, but the picture of Theon that has come down to us is one of Theon as a disseminator and conserver of knowledge, rather than an innovator. Moreover, Neugebauer has given us grounds to believe Ptolemy to have been the inventor.

Although Theon's work on the astrolabe is now regarded as lost, Neugebauer finds such similarities between later works that they must derive from a common source. This source he believes to be Theon. He further argues (because of the exact correspondence described above) that what Theon wrote was a Commentary on an earlier book by Ptolemy.

This gives us the background to Synesius' *De Dono Astrolabii*. Writing to Paionos, he states that he designed the astrolabe himself with help from Hypatia and had it crafted by the very best of silver-smiths. The inference is that the theory of the astrolabe and the details of its construction were passed down from Ptolemy, via Theon, to Hypatia, who in her turn taught Synesius.

11 The Hydroscope

Letter 15 of Synesius begins: "I am reduced to this, that I have to have a hydroscope." The letter then goes on to ask her to make him one, to quite detailed specifications. The question of what he needed is puzzling. The general presumption is that he was ill.

The term "hydroscope" usually implies a *clepsydra* or water-clock, but this seems inappropriate as a translation in this case. Why should he be, even if brought so low, in such urgent need of a water-clock? FitzGerald believes that Fermat (yes, *the* Fermat) [3] is right in suggesting that what Synesius needed was a hydrometer, that is to say, a densimeter. This makes much more sense of the specifications, which refer to the need to measure the *weight* of the water (the *clepsydra* measures the *volume*), and describe an instrument that sounds very like a hydrometer.

The suggestion is that Synesius needed it in his illness somehow to measure a medicine he was taking (or less plausibly the salinity of his drinking water). Hydrometers are now used, as they well may then have been, to measure the alcoholic contents of fermented or distilled liquors. Possibly Synesius was making his own medicine by some such means. My friend and colleague Charles Hunter (Department of Anatomy, Monash University) however offers a novel suggestion — that the "hydroscope" was in fact a urinometer and that the dosage of some diuretic was calculated by reference to the specific gravity of the urine.

12 Assessment

What we know of Hypatia is little enough; what we know of her Mathematics is only a small subset of that little. There is evidence that she was greatly regarded as a teacher and a scholar. The range of her acknowledged expertise was considerable. She edited works of Geometry, Algebra and Astronomy, knew how to make astrolabes and "hydrosopes",

and did a lot else besides. One cannot but be impressed with this breadth of interest. Moreover, at the time of her death (assuming with Toomer [28] that Theon pre-deceased her) she was in fact the greatest mathematician then living in the Greco-Roman world, very likely the world as a whole.

She is variously described as a philosopher, a teacher of Philosophy, a mathematician and astronomer, a learned woman and a geometer.

We can understand the term “philosopher” in two senses: it has the technical sense that it retains to this day, but it also has a generic meaning of “thinker”. Theon also is described in the sources as a philosopher. But this is surely in the second sense; Theon clearly emerges as a specialist mathematician and astronomer — the *Suda* goes on to say as much. Hypatia does not (unless one accords weight to the quote in Section 6 above); the *Suda* is at some pains to make this clear. “She also took up other [non-mathematical] branches of philosophy and though a woman she cast [an academic robe] around herself and appeared in the centre of the city” (Rist’s translation) — the *Suda* then proceeds to describe the Philosophy she taught, mentioning the work of Plato and Aristotle, in particular.

However, if we restrict consideration to Mathematics alone, we may well query the usual judgment that Hypatia outclassed her father. It comes from the PG and modern sources regularly repeat it uncritically. We may also deduce it from Theon’s heading to his *Commentary* of Book 3 of the *Almagest*.

We may still however dispute this opinion and indeed argue the opposite. That a fond father might recognise and promote his daughter’s improvement of one of his own works is understandable enough. That ecclesiastical historians, of whom we have no evidence of mathematical ability, might use fame or even notoriety as an index of talent is equally so. But this does not end the matter.

While it is of course too much to posit a universal theory of natural selection of scholarly works (it being by no means *always* true that the best works are the survivors) nonetheless scholars of earlier times preserved, translated and taught from those works they adjudged as valuable, much as we do today. In fact, we do know something of the principle of natural selection that operated. Because the focus had moved from research to conservation, those works were preserved that were well regarded as *textbooks* [29]. Many research works from the period are lost.

We have no evidence of research Mathematics on the part of either father or daughter. What we can

reconstruct of their Mathematics suggests to us that they edited, preserved, taught from and supplied minor addenda to the works of others. A great deal of Theon’s work survives and at most a small part of Hypatia’s. In other words Theon was seen as the better text-writer, even if he himself generously demurred in one case.

Where Hypatia *does* quite clearly outshine Theon is in her reputation as a teacher. She was revered as such and no similar endorsement of Theon has come down to us. (It is perfectly possible that this is the basis of the original statement.) We are left with a well-attested account of a popular, charismatic and versatile teacher. And that, I suggest, is the best picture we can form of her.

Addendum: Too late for mention in the main article, I was made aware of the lengthy discussion of Hypatia by W. R. Knorr [*Textual Studies in Ancient and Medieval Geometry* (Boston: Birkhauser, 1989)]. Beginning from a stylistic analysis of Book Three of Theon’s *Commentary* on the *Almagest*, Knorr builds an elaborate and detailed, though speculative, argument to attribute several other lost works to Hypatia. In particular, he suggests that Eutocius’ *Commentary* on Apollonius’ *Conics* in fact derives from Hypatia’s earlier *Commentary*, the one mentioned in the *Suda*.

References

1. Deakin, M. A. B., *Hypatia the Mathematician, Monash University History of Mathematics Pamphlet 52* (1991).
2. Delambre, J. B. J., *Histoire de l’Astronomie Ancienne 2* (New York: Johnson, 1965; reprint of an 1817 original), 317.
3. Fermat, P., *Oeuvres 1* (Ed. P. Tannery and C. Henry) (Paris: Gauthier, 1891), 352–365.
4. FitzGerald, A. (ed. and trans.), *The Letters of Synesius of Cyrene* (London: Oxford University Press, 1926).
5. Gibbon, E., *The Decline and Fall of the Roman Empire* (first published 1776–1788; many subsequent editions), Chapter 47.
6. Gow, J. *A Short History of Greek Mathematics* (New York: Chelsea, 1968; reprint of an 1884 original), 312–313.
7. Heath, T. L., *Diophantus of Alexandria* (Cambridge University Press, 1885; Dover reprint, 1964).
8. —, Review of Ref. [21], *Class. Rev.* 52 (1938), 40.

9. Hubbard, E., *Little Journeys to the Homes of Great Teachers* 23 (4) (East Aurora, NY: Roycrofters, 1908).
10. Kingsley, C., *Hypatia* (first published 1851; many subsequent editions).
11. Kramer, E. E., Article on Hypatia, *DSB* 6, 615–616.
12. Marrou, H. I., Synesius of Cyrene and Alexandrian Neoplatonism, in *The Conflict between Paganism and Christianity in the Fourth Century* (Ed. A. Momigliano) (Oxford: Clarendon, 1963), 126–150.
13. Migne, J.-P. (ed.), *Patrologiae Graecae* (Paris: Migne, 1857–1866).
14. Mueller, I., Hypatia, in *Women of Mathematics: A Biobibliographic Sourcebook* (Ed. L. S. Grinstein and P. J. Campbell) (New York: Greenwood, 1987).
15. Neugebauer, O., The Early History of the Astrolabe, *Isis* 40 (1949), 240–256.
16. —, *A History of Ancient Mathematical Astronomy* (Berlin: Springer, 1975).
17. North, J. D., The Astrolabe, *Sci. Am.* 230 (1), (Jan. 1974), 96–106.
18. Rashed, R., Review of Ref. [24], *Math. Rev.* 85:01006.
19. Richeson, A. W., Hypatia of Alexandria, *Nat. Math. Mag.* 15 (1940), 74–82.
20. Rist, J. M., Hypatia, *Phoenix* 19 (1965), 214–225.
21. Rome, A., *Commentaires de Pappus et de Théon d'Alexandrie sur l'Almageste* (2), *Studi e Testi* (Vatican) 72 (1936).
22. —, *Commentaires de Pappus et de Théon d'Alexandrie sur l'Almageste* (3), *Studi e Testi* (Vatican) 106 (1943).
23. —, Le troisième livre des commentaires sur l'“Almageste” par Théon et Hypatie, (Paris: Presses universitaires de France, 1926; excerpted from *Am. Soc. Sci. Bruxelles* 46, 1926).
24. Sesiano, J., *Books IV to VII of Diophantus' Arithmetica* (New York: Springer, 1982).
25. Tannery, P., L'article de Suidas sur Hypatia, *Ann. Fac. Lettres Bordeaux* 2 (1880), 197–200.
26. —, *Diophanti Alexandrini opera omnia* (Leipzig: Teubner, 1893–1895).
27. Toomer, G. J., Article on Apollonius, *DSB* 1, 179–193.
28. —, Article on Theon, *DSB* 13, 321–325.
29. —, Lost Greek Mathematical Works in Arabic Translation, *Math. Intell.* 6 (2) (1984), 32–38.

The Evolution of Mathematics in Ancient China

FRANK SWETZ

Mathematics Magazine 52 (1979), 10–19

A popular survey book on the development of mathematics has its text prefaced by the following remarks:

Only a few ancient civilizations, Egypt, Babylon, India and China, possessed what may be called the rudiments of mathematics. The history of mathematics and indeed the history of western civilization begins with what occurred in the first of these civilizations. The role of India will emerge later, whereas that of China may be ignored because it was not extensive and moreover has no influence on the subsequent development of mathematics [1].

Even most contemporary works on the history of mathematics reinforce this impression, either by neglecting or depreciating Chinese contributions to the development of mathematics [2]. Whether by ignorance or design, such omissions limit the perspective one might obtain concerning both the evolution of mathematical ideas and the place of mathematics in early societies. In remedying this situation, western historians of mathematics may well take heed of Whittier's admonition [3]:

We lack but open eye and ear
To find the Orient's marvels here.

Language barriers may limit this quest for information; however, a search of English language sources will reveal that there are many "marvels" in Chinese mathematics to be considered.

1 Legend and fact

The origins of mathematical activity in early China are clouded by mysticism and legend. Mythological Emperor Yü is credited with receiving a divine gift

from a Lo river tortoise. The gift in the form of a diagram called the *Lo shu* is believed to contain the principles of Chinese mathematics, and pictures of Yü's reception of the *Lo shu* have adorned Chinese mathematics books for centuries. This fantasy in itself provides some valuable impressions about early Chinese science and mathematics. Yü was the patron of hydraulic engineers; his mission was to control the flood-prone waters of China and provide a safe setting in which a water-dependent civilization could flourish. The users of science and mathematics in China were initially involved with hydraulic engineering projects, the construction of dikes, canals, etc., and with the mundane tasks of logistically supporting such projects. A close inspection of the contents of the *Lo shu* reveals a number configuration (Figure 1) which would be known later in the West as a magic square. For Chinese soothsayers and geomancers from the Warring State period of Chinese history (403–221 B.C.) onward, this square, comprised of numbers, possessed real magical qualities because in it they saw a plan of universal harmony based on a cosmology predicated on the dualistic theory of the *Yin* and the *Yang* [4].

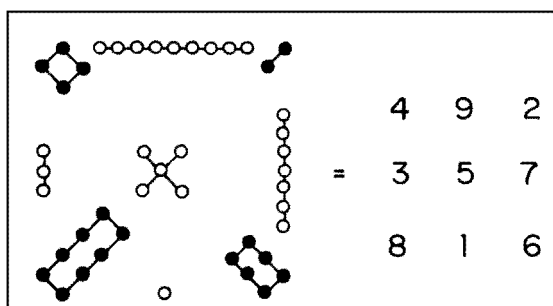


Figure 1.

When stripped of ritualistic significance, the principles used in constructing this first known magic square are quite simple and can best be described by use of diagrams as shown as follows:

1. Construct a natural square.

7	4	1
8	5	2
9	6	3

2. Distort it into a diamond.

		1		
	4		2	
7		5		3
	8		6	
		9		

3. Exchange corner elements.

		9		
	4		2	
3		5		7
	8		6	
		1		

4. Compress back into a square.

4	9	2
3	5	7
8	1	6

The construction and manipulation of magic squares became an art in China even before the concept was known in the West [5]. Variations of the *Lo shu* technique were used in constructing magic squares of higher order with perhaps the most impressive square being that of order nine:

1. Start with a natural square

1	10	19	28	37	46	55	64	73
2	11							.
3	12							.
4	13							.
5	14							.
6	15							.
7	16							.
8	17							.
9	18							81

2. Then fold each row into a square of order 3 (example using row 1)

55	28	1
64	37	10
73	46	19

3. Apply the *Lo shu* technique

28	73	10
19	37	55
64	1	46

4. The nine resulting magic squares of order 3 are then positionally ordered according to the correspondence of the central element in their bottom rows with the numbers of the *Lo shu*:

4,9,2; 3,5,7; 8,1,6.

31	76	13	36	81	18	29	74	11
22	40	58	27	45	63	20	38	56
67	4	49	72	9	54	65	2	47
30	75	12	32	77	14	34	79	16
21	39	57	23	41	59	25	43	61
66	3	48	68	5	50	70	7	52
35	80	17	28	73	10	33	78	15
26	44	62	19	37	55	24	42	60
71	8	53	64	1	46	69	6	51

While the *Lo shu* provides some intriguing insights into early mathematical thinking, its significance in terms of potential scientific or technological achievement is negligible. Historically, the first true evidence of mathematical activity can be found in numeration symbols on oracle bones dated from the Shang dynasty (14th century B.C.). Their numerical inscriptions contain both tally and code symbols, are clearly decimal in their conception, and employ a positional value system. The Shang numerals for the numbers one through nine are illustrated in Figure 2.



Figure 2.

By the time of the Han Dynasty (2nd century B.C.–4th century A.D.), the system had evolved into a codified notation that lent itself to computational algorithms carried out with a counting board and set of rods. The numerals and their computing-rod configurations are illustrated in Figure 3.

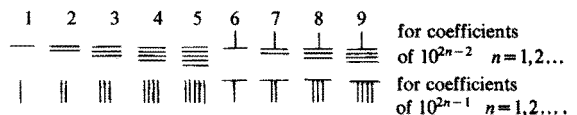


Figure 3.

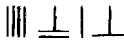


Figure 4.

Thus in this system 4716 would be represented as in Figure 4 [6]. Occasionally the symbol \times was used as an alternative symbol for 5, in the even power places.

Counting boards were divided into columns designating positional groupings by 10. The resulting facility with which the ancient computers could carry out algorithms attests to their full understanding of decimal numeration and computation. As an example, consider the counting board method of multiplying 2 three-digit numbers, as illustrated in Figure 5.

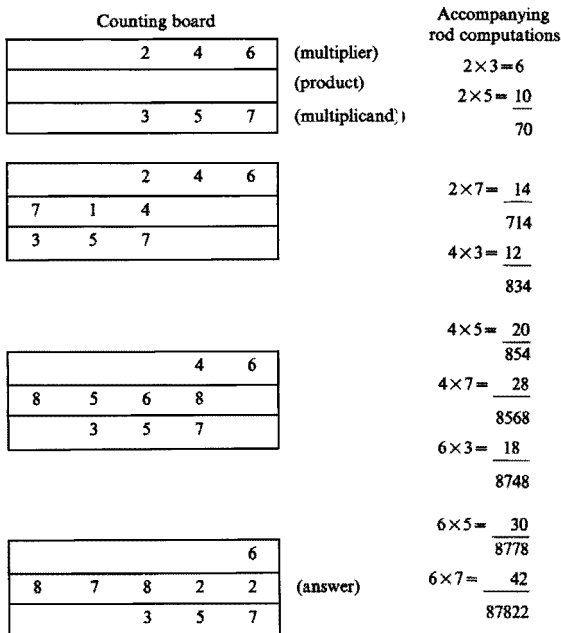


Figure 5.

The continual indexing of partial products to the right as one multiplies by smaller powers of ten testifies to a thorough understanding of decimal notation. In light of such evidence, it would seem that the Chinese were the first society to understand and efficiently utilize a decimal numeration system [7]. If one views a popular schematic of the evolution of our modern system of numeration (Figure 6) and places the Chinese system in the appropriate chronological position, an interesting hypothesis arises, namely that the numeration system commonly used in the modern world had its origins 34 centuries ago in Shang China!

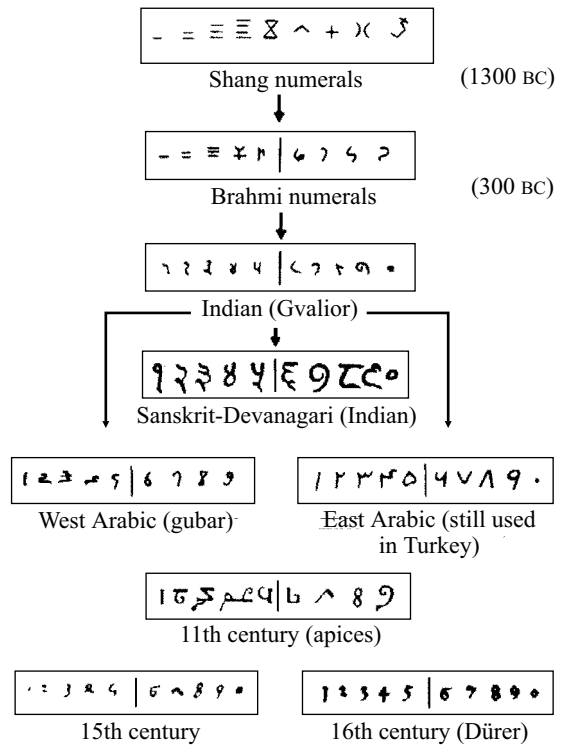


Figure 6.

2 The systematization of early Chinese mathematics

The oldest extant Chinese text containing formal mathematical theories is the *Arithmetic Classic of the Gnomon and the Circular Paths of Heaven*, [Chou pei suan ching]. Its contents date before the third century B.C. and reveal that mathematicians of the time could perform basic operations with fractions according to modern principles employing the concept of common denominator. They were knowledgeable in the principles of an empirical geometry and made use of the “Pythagorean theorem”. A diagram (see Figure 7) in the *Chou pei* presents the oldest known demonstration of the validity of this theorem. This diagram, called the *hsuan-thu* in Chinese, illustrates the arithmetic-geometric methodology that predominates in early Chinese mathematical thinking and shows how arithmetic and geometry could be merged to develop algebraic processes and procedures. If the oblique square of the *hsuan-thu* is dissected and the pieces rearranged so that two of the four congruent right triangles are joined with the remaining two to form two rectangles, then the resulting figure comprised of two rectangles and

弦圖

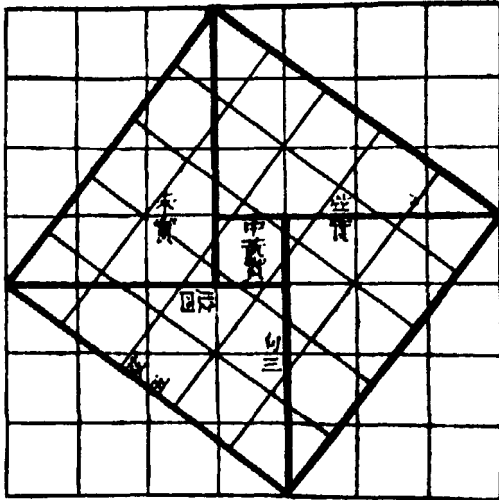


Figure 7.

one small square have the same area as their parent square. Further, since the new configuration can also be viewed as being comprised of two squares whose sides are the legs of the right triangles, this figure demonstrates that the sum of the squares of the legs of a right triangle is equal to the square of the hypotenuse [8]. The processs involved in this intuitive, geometric approach to obtain algebraic results was called *chi-chü* or “the piling up of squares” [9].

The next historical text known to us is also a Han work of about the third century B.C. It is the *Nine Chapters on the Mathematical Art* [*Chiu chang suan shu*], and its influence on oriental mathematics may be likened to that of Euclid’s *Elements* on western mathematical thought. The *Chiu chang*’s chapters bear such titles as surveying of land, consultations on engineering works, and impartial taxation, and confirm the impression that the Chinese mathematics of this period centered on the engineering and bureaucratic needs of the state. Two hundred and forty-six problem situations are considered, revealing in their contents the fact that the Chinese had accumulated a variety of formulas for determining the areas and volumes of basic geometric shapes. Linear equations in one unknown were solved by a rule of false position. Systems of equations in two or three unknowns were solved simultaneously by computing board techniques that are strikingly similar to mod-

ern matrix methods. While algebraists of the ancient world such as Diophantus or Brahmagupta used various criteria to distinguish between the variables in a linear equation [10], the Chinese relied on the organizational proficiency of their counting board to assist them in this chore. Using a counting board to work a system of equations allowed the Chinese to easily distinguish between different variables.

Consider the following problem from the *Chiu chang* and the counting board approach to its solution.

Of three classes of cereal plants, 3 bundles of the first, 2 of the second and 1 of the third will produce 39 *tou* of corn after threshing; 2 bundles of the first, 3 of the second and 1 of the third will produce 34 *tou*; while 1 of the first, 2 of the second and 3 of the third will produce 26 *tou*. Find the measure of corn contained in one bundle of each class [11]. [1 *tou* = 10.3 liters]

This problem would be set up on the counting board as:

1	2	3	1st class grain
2	3	2	2nd class grain
3	1	1	3rd class grain
26	34	39	Number of <i>tou</i>

Using familiar notation this matrix of numbers is equivalent to the set of equations

$$\begin{aligned} 3x + 2y + z &= 39 \\ 2x + 3y + z &= 34 \\ x + 2y + 3z &= 26 \end{aligned}$$

which are reduced in their tabular form by appropriate multiplications and subtraction to

$$\begin{aligned} 3x + 2y + z &= 30 \\ &36y &= 153 \\ &&36z &= 99 \end{aligned}$$

and

$$\begin{aligned} 36x &= 333 \\ 36y &= 153 \\ 36z &= 99. \end{aligned}$$

Thus $x = 333/36, y = 153/36$ and $z = 99/36$.

A companion problem from the *Chiu chang* involves payment for livestock and results in the system of simultaneous equations:

$$\begin{aligned} -2x + 5y - 13z &= 1000 \\ 3x - 9y + 3z &= 0 \\ -5x + 6y + 8z &= -600. \end{aligned}$$

Rules provided for the solution treat the addition and subtraction of negative numbers in a modern fashion; however, procedures for the multiplication and division of negative numbers are not found in a Chinese work until the Sung dynasty (+1299). Negative numbers were represented in the computing scheme by the use of red rods, while black computing rods represented positive numbers. Zero was indicated by a blank space on the counting board. This evidence qualifies the Chinese as being the first society known to use negative numbers in mathematical calculations.

The *Chou pei* contains an accurate process of extracting square roots of numbers. The ancient Chinese did not consider root extraction a separate process of mathematics but rather merely a form of division [12]. Let us examine the algorithm for division and its square root variant. The division algorithm is illustrated in Figure 8 for the problem $166536 \div 648$.

The Chinese technique of root extraction depends on the algebraic proposition

$$(a + b + c)^2 = a^2 + 2ab + b^2 + 2(a + b)c + c^2 = a^2 + (2a+b)b + (2[a+b]+c)c \quad (1)$$

which is geometrically substantiated by the diagram given in Figure 9. This proposition is incorporated directly into a form of division where $\sqrt{N} = a + b + c$. The counting board process for extracting the square root of 55225 is briefly outlined in Figure 10. Root extraction was not limited to three digit results, for the Chinese were able to continue the process to several decimal places as needed. Decimal fractions were known and used in China as far back as the

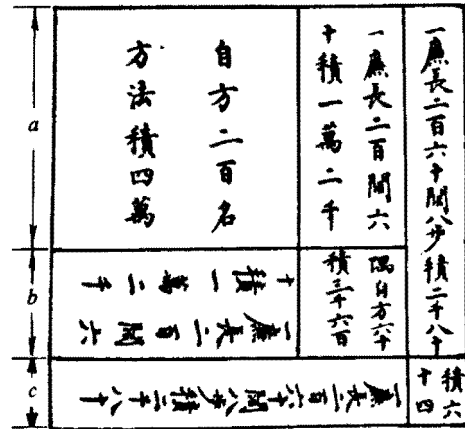


Figure 9. A geometric proof of Equation (1)

Counting board layout	Accompanying rod computations	Explanations
2 (quotient)	166500 - 120000 = (200 × 600)	200 is chosen as the first partial quotient
166536 (dividend)	46500 - 8000 = (200 × 40)	
648 (divisor)	38500 - 1600 = (200 × 8) 36900	
25	36930 - 30000 = (50 × 600)	50 is chosen as the second partial quotient
36936	6930 - 2000 = (50 × 40)	
648	4930 400 = (50 × 8) 4530	
257	4536 - 4200 = (7 × 600)	7 is chosen as the third partial quotient
4536	336 - 280 = (7 × 40)	
648	56 - 56 = (7 × 8) 0	process is finished

Figure 8. $166536 \div 648$

Algebraic Significance Numerical entries on board

N 1	55225 1
a $N - a^2$ $a \times 10000$ 10000	2 15225 20000 10000
$a + b$ $N - a^2$ $(2a + b)b \times 100$ $(2a + b) \times 100$ 100	23 15225 12900 4300 100
$a + b$ $N - [a^2 + (2a + b)b]$ $(2a + b) \times 100$ 100	23 2325 4300 100
$a + b + c$ $N - (a + b)^2 - [2a(a + b) + c]c$ $2(a + b) + c$	235 0 465

Figure 10. The calculation of $\sqrt{55225} = 235$. The 1 in the upper box represents an indexing rod that determines the decimal value of the divisors used. At the beginning of the process, it is moved to the left in jumps of two decimal places until it establishes the largest power of ten that can be divided into the designated number. After each successful division, the rod is indexed two positional places to the right.

5th century B.C. Where a root was to be extracted to several decimal places, the computers achieved greater accuracy by use of the formulae [13]

$$\sqrt[n]{m} = \frac{\sqrt[n]{m10^{kn}}}{10^k}.$$

Cube root extraction was conceived on a similar geometric-algebraic basis and performed with equal facility.

Historians of mathematics often devote special consideration to the results obtained by ancient societies in determining a numerical value for π as they believe that the degree of accuracy achieved supplies a comparative measure for gauging the level of mathematical skill present in the society. On the basis of such comparisons, the ancient Chinese were far superior to their contemporaries in computational mathematical ability. Aided by a number system that included the decimalization of fractions and the possession of an accurate root extraction process the Chinese had obtained by the first century a value of π of 3.15147. The scholar Liu Hui in a third century commentary on the *Chiu chang* employed a “cutting of the circle method” — determining the area of a circle with known radius by polygonal approximations — to determine π as 3.141024. A successor, Tsu Chung-chih, refined the method in the fifth century to derive the value of π as 355/113 or 3.1415929 [14]. This accuracy was not to be arrived at in Europe until the 16th century.

3 Trends in Chinese algebraic thought

While the Chinese computational ability was indeed impressive for the times, their greatest accomplishments and contributions to the history of mathematics lay in algebra. During the Han period, the square and cube root extraction processes were being built upon to obtain methods for solving quadratic and other higher order numerical equations. The strategy for extending the square root process to solve quadratic equations was based on the following line of reasoning. If $x^2 = 289$, 10 would be chosen as a first entry approximation to the root, then

$$289 - (10)^2 = 189.$$

Let the second entry of the root be represented by y ; thus, $x = 10 + y$ or $(10 + y)^2 = 289$ which, if expanded, gives the quadratic equation

$$y^2 + 20y - 189 = 0.$$

By proceeding to find the second entry of the square root of 289, 7, we obtain the positive root for the quadratic $y^2 + 20y - 189 = 0$ [15].

By the time of the Sung Dynasty in the 13th century, mathematicians were applying their craft to solve such challenging problems as:

This is a round town of which we do not know the circumference or diameter. There are four gates (in the wall). Three *li* from the northern (gate) is a high tree. When we go outside of the southern gate and turn east, we must walk 9 *li* before we see the tree. Find the circumference and the diameter of the town. [1 *li* = 0.644 kilometers]

If the diameter of the town is allowed to be represented by x^2 , the distance of the tree from the northern gate, a , and the distance walked eastward, b , the following equation results.

$$x^{10} + 5ax^8 + 8a^2x^6 - 4a(b^2 - a^2)x^4 - 16a^2b^2x^2 - 16a^3b^2 = 0.$$

For the particular case cited above, the equation becomes

$$x^{10} + 15x^8 + 72x^6 - 864x^4 - 11604x^2 - 34992 = 0.$$

Sung algebraists found the diameter of the town to be 9 *li* [16].

The earliest recorded instance of work with indeterminate equations in China can be found in a problem situation of the *Chiu chang* where a system of four equations in five unknowns results [17]. A particular solution is supplied. A problem in the third century *Mathematical Classic of Sun Tzu* [*Sun Tzu suan ching*] concerns linear congruence and supplies a truer example of indeterminate analysis.

We have things of which we do not know the number; if we count by threes, the remainder is 2; if we count by fives, the remainder is 3; if we count by sevens, the remainder is 2. How many things are there? [18]

In modern form, the problem would be represented as:

$$N \equiv 2 \pmod{3} \equiv 3 \pmod{5} \equiv 2 \pmod{7}.$$

Sun’s solution is given by the expression

$$(70 \times 2) + (21 \times 3) + (15 \times 2) \equiv 23 \pmod{105}$$

which when analysed gives us the first application of the Chinese Remainder Theorem.

If m_1, \dots, m_k are relatively prime in pairs, there exist integers x for which simultaneously

$$x \equiv a_1 \pmod{m_1}, \dots, x \equiv a_k \pmod{m_k}.$$

All such integers x are congruent modulo $m = m_1 m_2 \dots m_k$. The existence of the Chinese Remainder Theorem was communicated to the west by Alexander Wylie, an English translator and mathematician in the employ of the nineteenth century Chinese court. Wylie recorded his findings in a series of articles, "Jottings on the Science of the Chinese; Arithmetic" which appeared in the *North China Herald* (Aug.-Nov.) 1852. The validity of the theorem was questioned until it was recognized as a variant of a formula developed by Gauss [19].

Perhaps the most famous Chinese problem in indeterminate analysis, in the sense of its transmission to other societies, was the problem of the "hundred fowls" (ca. 468).

A cock is worth 5 *ch'ien*, a hen 3 *ch'ien*, and 3 chicks 1 *ch'ien*. With 100 *ch'ien* we buy 100 fowls. How many cocks, hens, and chicks are there? [*ch'ien*, a small copper coin]

The development of algebra reached its peak during the later part of the Sung and the early part of the following Yuan dynasty (13th and 14th centuries). Work with indeterminate equations and higher order numerical equations was perfected. Solutions of systems of equations were found by using methods that approximate an application of determinants, but it wasn't until 1683 that the Japanese Seki Kowa, building upon Chinese theories, developed a true concept of determinants.

Work with higher numerical equations is facilitated by a knowledge of the binomial theorem. The testimony of the Chiu chang indicates that its early authors were familiar with the binomial expansion $(a + b)^3$, but Chinese knowledge of this theorem is truly confirmed by a diagram (Figure 11) appearing in the 13th century text *Detailed Analysis of the Mathematical Rules in the Nine Chapters*. [*Hsiang chieh chiu chang suan fa*.] It seems that "Pascal's Triangle" was known in China long before Pascal was even born.

While mathematical activity continued in the post-Sung period, its contributions were minor as compared with those that had come before. By the time of the Ming emperors in the 17th century, western mathematical influence was finding its way into China and the period of indigenous mathematical accomplishment had come to an end.

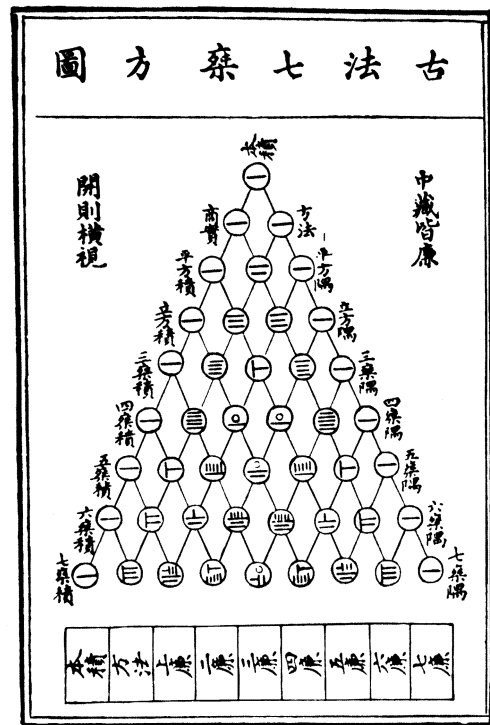


Figure 11.

4 Conclusions

Thus, if comparisons must be made among the societies of the pre-Christian world, the quality of China's mathematical accomplishments stands in contention with those of Greece and Babylonia, and during the period designated in the West as pre-Renaissance, the sequence and scope of mathematical concepts and techniques originating in China far exceeds that of any other contemporary society. The impact of this knowledge on the subsequent development of western mathematical thought is an issue that should not be ignored and can only be resolved by further research.

In part, such research will have to explore the strength and vitality of Arabic-Hindu avenues of transmission of Chinese knowledge westward. The fact that western mathematical traditions are ostensibly based on the logico-deductive foundations of early Greek thought should not detract from considering the merits of the inductively-conceived mathematics of the Chinese. After all, deductive systemization is a luxury afforded only after inductive and empirical experimentation has established a foundation from which theoretical considerations can proceed. Mathematics, in its primary state, is a tool

for societal survival; once that survival is assured, the discipline can then become more of an intellectual and aesthetic pursuit. Unfortunately, this second stage of mathematical development never occurred in China. This phenomenon — the fact that mathematics in China, although developed to a high art, was never elevated further to the status of an abstract deductive science — is yet another fascinating aspect of Chinese mathematics waiting to be explained.

Notes

1. Morris Kline, *Mathematics: A Cultural Approach* (Reading, Mass.: Addison-Wesley Publishing Co. 1962) p. 12.
2. In his 712-page *A History of Mathematics* (New York: John Wiley & Sons Inc., 1968) Carl Boyer devotes 12 pages to Chinese contributions; the latest revised edition of Howard Eves, *An Introduction to the History of Mathematics*, (New York: Holt, Rinehart and Winston, 1976) contains 6 pages on the history of Chinese mathematics. The contents of these pages are based on information given in an article by D. J. Struik, On Ancient Chinese Mathematics, *The Mathematics Teacher* 56 (1963), 424–432 and represent little of Eves' own research.
3. John Greenleaf Whittier, "The Chapel of the Hermits".
4. Under this system, the universe is ruled by Heaven through means of a process called the *Tao* ("the Universal way"). Heaven acting through the *Tao* expresses itself in the interaction of two primal forces, the *Yin* and the *Yang*. The *Yang*, or male force, was a source of heat, light and dynamic vitality and was associated with the sun; in contrast, the *Yin*, or female force, flourished in darkness, cold and quiet inactivity and was associated with the moon. In conjunction, these two forces influenced all things and were present individually or together in all physical objects and situations. In the case of numbers, odd numbers were *Yang* and even, *Yin*. For a harmonious state of being to exist, *Yin-Yang* forces had to be balanced.
5. For a fuller discussion of Chinese magic squares, see Schyler Camman, Old Chinese Magic Squares, *Sinologica* 7 (1962), 14–53; Frank Swetz, Mysticism and Magic in the Number Squares of Old China, *The Mathematics Teacher* 71 (January, 1978), 50–56.
6. The evolution of counting rod numerals continued for about 3000 years in China, i.e., 14th century BC–13th century AD. For a discussion of this process, see Joseph Needham, *Science and Civilization in China* (Cambridge: Cambridge University Press, 1955) vol. 3. pp. 5–17.
7. A strong case for this theory has been made by Wang Ling, The Chinese Origin of the Decimal Place-Value System in the Notation of Numbers. Communication to the 23rd International Congress of Orientalists, Cambridge, 1954.
8. Although a 3, 4, 5 right triangle is used in the demonstration, the Chinese generalized their conclusion for all right triangles. The 3, 4, 5 triangle was merely a didactical aid.
9. See Frank Swetz, The 'Piling Up of Squares' in Ancient China, *The Mathematics Teacher* 70 (1977), 72–79.
10. Diophantus (275 AD) spoke of unknowns of the first number, second number, etc., whereas Brahmagupta (628 AD) used different colors in written computations to distinguish between variables.
11. *Chiu chang suan shu*, Fang Cheng (chapter 8), problem 1.
12. For a discussion of the Chinese ability at root extraction, see Wang Ling and Joseph Needham, Horner's Method in Chinese Mathematics: Its Origins in the Root Extraction Procedures of the Han Dynasty, *T'oung Pao* 43 (1955), 345–88; Lam Lay Yong, The Geometrical Basis of the Ancient Chinese Square-Root Method, *Isis* (Fall, 1970), pp. 92–101.
13. A lengthy discussion of the use of this formula in Europe is given in D. E. Smith, *History of Mathematics* (New York: Dover Publishing Co., 1958 reprint) vol. II, p. 236.
14. The evolution of π in China is traced out in Lee Kiong-Pong, "Development of π in China", *Bulletin of the Malaysian Mathematical Society* 6 (1975), 40–47.
15. An actual computational procedure used in solving quadratics can be found in Ho Peng Yoke, The Lost Problems of the Chang Ch'iu-chien Sua Ching, a Fifth Century Chinese Mathematical Manual, *Oriens Extremus* (1965), 12.
16. For a detailed discussion of the solution of this problem see Ulrich Libbrecht, *Chinese Mathematics in the Thirteenth Century* (Cambridge, Mass.: The MIT Press, 1973) pp. 134–40.
17. *Chiu chang suan shu*, chapter 8, problem 13:
There is a common well belonging to five families; (if we take) 2 lengths of rope of family *X*, the remaining part equals 1 length of rope of family *Y*; the remaining part from 3 ropes of *Y* equals 1 rope of *Z*; the remaining part from 4 ropes of *Z* equals 1 rope of *V*; the lacking part remaining from 5 ropes of *V* equals 1 rope of *U*; the remaining part from 6 ropes of *U* equals 1 rope of *X*. In all instances if one gets the missing length of rope, the combined lengths will reach (the water). Find the depth of the well and the length of the ropes.

If we let W equal the depth of the well, the following system of equations result:

$$2X + Y = W$$

$$3Y + Z = W$$

$$4Z + V = W$$

$$5V + U = W$$

$$6U + X = W$$

which are readily reduced to:

$$2X - 2Y - Z = 0$$

$$2X + Y - 4Z - V = 0$$

$$2X + Y - 5V - U = 0$$

$$X + Y - 6U = 0.$$

18. *Sun Tzu suan ching*, chapter 3, problem 10.
19. See the discussion of the Chinese Remainder Theorem in Oystein Ore, *Number Theory and its History* (New York: McGraw-Hill Inc., 1948) pp. 245–49.

Liu Hui and the First Golden Age of Chinese Mathematics

PHILIP D. STRAFFIN, JR.

Mathematics Magazine 71 (1998), 163–181

1 Introduction

Very little is known of the life of Liu Hui, except that he lived in the Kingdom of Wei in the third century A.D., when China was divided into three kingdoms at continual war with one another. What is known is that Liu was a mathematician of great power and creativity. Liu's ideas are preserved in two works which survived and became classics in Chinese mathematics. The most important of these is his commentary, dated 263 A.D., on the *Jiuzhang suanshu*, the great problem book known in the West as the *Nine Chapters on the Mathematical Art*. The second is an independent work on mathematics for surveying, the *Haidao suanjing*, known as the *Sea Island Mathematical Manual*.

In this paper I would like to tell you about some of the remarkable results and methods in these two works. I think they should be more widely known, for several reasons. First, we and our students should know more about mathematics in other cultures, and we are probably less familiar with Chinese mathematics than with the Greek, Indian and Islamic traditions more directly linked to the historical development of modern mathematics. Second, Western mathematicians who do know something about the Chinese tradition often characterize Chinese mathematics as calculational and utilitarian rather than theoretical. Chinese mathematicians, it is said, developed clever methods, but did not care about mathematical justification of those methods. For example,

Mathematics was overwhelmingly concerned with practical matters that were important to a bureaucratic government: land measurement and surveying, taxation, the making of

canals and dikes, granary dimensions, and so on. . . Little mathematics was undertaken for its own sake in China. [2, p. 26]

While there is justice in this generalization, Liu Hui and his successors Zu Chongzhi and Zu Gengzhi were clearcut exceptions. Their methods were different from those of the Greeks, but they gave arguments of cogency and clarity which we can honor today, and some of those arguments involved infinite processes which we recognize as underlying the integral calculus.

My final reason is that I think mathematical genius should be honored wherever it is found. I hope you will agree that Liu Hui is deserving of our honor.

To understand the context of Liu's work, we must first consider the state of Chinese mathematical computation in the third century A.D. We will then look at the general nature of the *Nine Chapters* and Liu's commentary on it, and at Liu's *Sea Island Mathematical Manual*. I will then focus on three of Liu's most remarkable achievements in geometry — his calculation of π , his derivation of the volume of pyramidal solids, and his work on the volume of a sphere and its completion by Zu Gengzhi.

2 Chinese calculation in the first century A.D.

From at least the period of the Warring States (475–221 B.C.) a base ten positional number system was in common use in China [12]. Calculations were done using rods made from bone or bamboo, on a counting board marked off into squares. The numerals from 1 to 9 were represented by rods as in Figure 1.

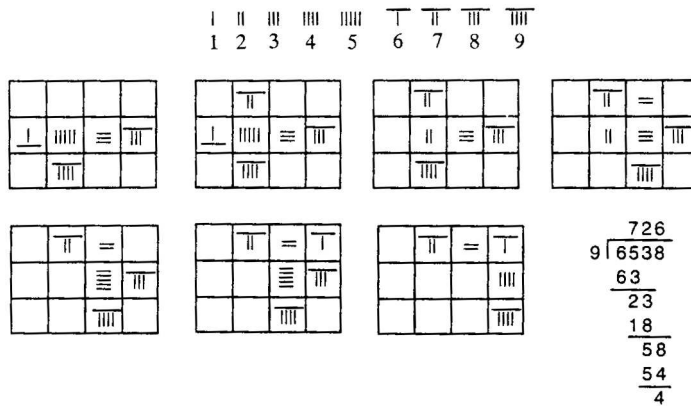


Figure 1. Numerals and the division algorithm

Their placement in squares, from left to right, represented decreasing powers of ten. Rods representing odd powers of ten were rotated 90° for clarity in distinguishing the powers. A zero was represented simply by a blank square, called a *kong*, where the marking into squares prevented the ambiguity sometimes present in, say, the Babylonian number system.

There were efficient algorithms for addition, subtraction, multiplication and division. For example, the division algorithm is shown in Figure 1, except that you should imagine the operations being done rapidly with actual sticks. Notice the close relationship to our modern long division algorithm, although subtraction is easier because sticks are physically removed. In fact, it is identical to the division algorithm given by al-Khwarizmi in the ninth century and later transmitted to Europe, raising the complicated problem of possible transmission through India to the West [12]. (See [17] for a conservative discussion.)

Notice how the answer $726\frac{4}{9}$ ends up with 726 in the top row, and then 4 above 9. This led Chinese calculators to represent fractions by placing the numerator above the denominator on the counting board. By the time of the *Nine Chapters* there was a completely developed arithmetic of fractions: they could be multiplied, divided, compared by cross multiplication, and reduced to lowest form using the “Euclidean algorithm” to find the largest common factor of the numerator and denominator. Addition was performed as $\frac{a}{b} + \frac{c}{d} = \frac{ad+bc}{bd}$, and then the fraction was reduced if necessary. In the *Nine Chapters*, 160 of the 246 problems involve computations with fractions [11].

We will see that Chapter Eight of the *Nine Chapters* solves systems of linear equations by the method

known in the West as “Gaussian Elimination” after C. F. Gauss (1777–1855), which, of course, involves subtracting one row of numbers from another. In the course of such calculations, it is inevitable that negative numbers will arise. This presented no problems to Chinese calculators: two colors of rods were used, and correct rules were given for manipulating the colors. Liu Hui suggested in his commentary on the *Nine Chapters* that negative numbers be treated abstractly:

When a number is said to be negative, it does not necessarily mean that there is a deficit. Similarly, a positive number does not necessarily mean that there is a gain. Therefore, even though there are red (positive) and black (negative) numerals in each column, a change in their colors resulting from the operations will not jeopardize the calculation. [17, pp. 201–202]

Perhaps most remarkably, Chinese mathematicians had developed by the time of the *Nine Chapters* efficient algorithms for computing square roots and cube roots of arbitrarily large numbers. The algorithm for the square root computed the root digit by digit, by the same method which used to be taught in American schools before the coming of the calculator. Martzloff [17] works through an example, and Lam [11] shows how it would look on a counting board. The algorithm for finding cube roots was similar, although, of course, more complicated.

In other words, by the time of the *Nine Chapters* the Chinese had developed a number system and a collection of calculational algorithms essentially equivalent to our modern system, with the exception of decimal fractions.

3 Nine Chapters on the Mathematical Art

Nine Chapters on the Mathematical Art is a compilation of 246 mathematical problems loosely grouped in nine chapters. Some of its material predates the great book-burning and burial-alive of scholars of 213 B.C., ordered by emperor Shih Huang-ti of the Qin dynasty. Indeed, Liu Hui writes in the preface of his commentary

In the past, the tyrant Qin burnt written documents, which led to the destruction of classical knowledge . . . Because of the state of deterioration of the ancient texts, Zhang Cang and his team produced a new version . . . filling in what was missing. [17, p. 129]

It is believed that the *Nine Chapters* were put in their final form sometime before 100 A.D. It “became, in the Chinese tradition, the mandatory reference, the classic of classics.” [17, p. 14] At the time of this writing there is no complete English translation of the *Nine Chapters*, although there are many scholarly Chinese editions, and translations into Japanese, German and Russian. An English translation by J. N. Crossley and Shen Kangsheng is in preparation, to be published by Springer Verlag. For summaries, see [11], [17], [18], [21].

The format of the *Nine Chapters* is terse: a problem, its answer, and a recipe for obtaining the answer. Usually no justification is given for the method of solution. Just the facts.

Chapter One has many problems on the arithmetic of fractions, and a section on computing areas of planar figures, with correct formulas for rectangles, triangles and trapezoids. Here’s a problem on the area of a circle:

1.32: There is a circular field, circumference 181 *bu* and diameter $60\frac{1}{3}$ *bu*. Find the area of the field.

Answer: 11 *mu* $90\frac{1}{12}$ *bu*. (1 *mu* = 240 *bu*)

Method: Mutually multiply half of the circumference and half of the diameter to obtain the area in *bu*. Or multiply the diameter by itself, then by 3 and divide by 4. Or multiply the circumference by itself and divide by 12. [11, p.13]

The first method is correct, but the data of the problem and the other two methods assume that the ratio of the circumference of a circle to its diameter, which we call π , is three. This assumption is made throughout the *Nine Chapters*.

Chapter Two is a series of commodity exchange problems involving proportions. Chapter Three concerns problems of “fair division.” The solutions given may not seem very fair to us:

3.8: There are five persons: Dai Fu, Bu Geng, Zan Niao, Shang Zao and Gong Shi. They pay a total of 100 *qian*. A command desired that the highest rank pays the least, and the successive ones gradually more. Find the amount each has to pay.

Answer: Dai Fu pays $8\frac{104}{137}$ *qian*; Bu Geng pays $10\frac{130}{137}$ *qian*; Zan Niao pays $14\frac{82}{137}$ *qian*; Shang Zao pays $21\frac{123}{137}$ *qian*; Gong Shi pays $43\frac{109}{137}$ *qian*. [11, p. 21]

The method calls for dividing the cost in proportions $\frac{1}{5} : \frac{1}{4} : \frac{1}{3} : \frac{1}{2} : 1$, which gives practice in adding fractions, but badly exploits the lowest rank person!

Chapter Four contains problems asking for the calculation of square roots and cube roots. The last problem of Chapter Four is

4.24: There is a sphere of volume 16441866437500 *chi*. Find the diameter.

Answer: 14300 *chi*.

Method: Put down the volume in *chi*, multiply by 16 and divide by 9. Extract the cube root of the result to get the diameter of the sphere. [11, p. 23]

This gives the formula $V = \frac{9}{16}d^3$ for the volume of a sphere in terms of its diameter, which isn’t correct even if we take $\pi = 3$.

Chapter Five asks for the volumes of a number of solids, including several different kinds of pyramids, frustrums of pyramids, cones and their frustrums, and a wedge with a trapezoidal base. The given formulas are all correct, but no hint is given of how they were derived.

Chapter Six deals with fair division in a much more realistic way than the problems in Chapter Three. There are problems on transporting grain, taxation and irrigation. There are also some less realistic problems which make one wonder how Chinese students must have felt about “word problems”:

6.14: There is a rabbit which walks 100 *bu* before it is chased by a dog. When the dog has gone 250 *bu*, it stops and is 30 *bu* behind the rabbit. If the dog did not stop, find how many more *bu* it would have to go before it reaches the rabbit.

Answer: $107\frac{1}{7}$ *bu*. [11, p. 28]

Chapter Seven has a number of problems involving two linear equations in two unknowns, usually

solved by the method of “false position.” Problems in Chapter Eight involve solving n linear equations in n unknowns for n up to 5. The method of solution, described in detail, is Gaussian elimination on the appropriate matrix represented on the counting board. The Chinese called this method *fangcheng*. See [17] for an extended example. Perhaps the most interesting problem is

8.13: There are five families which share a well. 2 of A’s ropes are short of the well’s depth by 1 of B’s ropes. 3 of B’s ropes are short of the depth by 1 of C’s ropes. 4 of C’s ropes are short by 1 of D’s ropes. 5 of D’s ropes are short by 1 of E’s ropes. 6 of E’s ropes are short by 1 of A’s ropes. Find the depth of the well and the length of each rope.

Answer: The well is 721 *cun* deep. A’s rope is 265 *cun* long. B’s rope is 191 *cun* long. C’s rope is 148 *cun* long. D’s rope is 129 *cun* long. E’s rope is 76 *cun* long. [11, p. 37]

Notice that this problem involves five equations and six unknowns, and thus is indeterminate. Liu Hui pointed out that the solution gives only the necessary proportions for the lengths. It is also the smallest solution in integer lengths.

The problems in Chapter Nine involve right triangles and the “Pythagorean” theorem, which had long been independently known in China, where it was called the *gou-gu* theorem [26]. No proof is given of this theorem, or of a correct formula for the diameter of the inscribed circle in a right triangle. Similar right triangles are used to solve surveying problems involving one unknown distance or length.

4 Liu Hui’s commentary

The *Nine Chapters* presents its solution methods without justification. Liu Hui in his commentary set himself the goal of justifying those methods. One reason was practical, as Liu wrote about the *Nine Chapters*’s use of 3 for the ratio of the circumference of a circle to its diameter:

Those who transmit this method of calculation to the next generation never bother to examine it thoroughly but merely repeat what they learned from their predecessors, thus passing on the error. Without a clear explanation and definite justification it is very difficult to separate truth from fallacy. [20, p. 349]

Another reason has to do with seeing and appreciating the logical structure of mathematics:

Things are related to each other through logical reasons so that like branches of a tree, diversified as they are, they nevertheless come out of a single trunk. If we elucidate by prose and illustrate by pictures, then we may be able to attain conciseness as well as comprehensiveness, clarity as well as rigor. [20, p. 355]

In this section, we’ll begin our examination of Liu’s attempt to attain “clarity as well as rigor” by looking at five of his contributions.

Problems in Chapter Four of the *Nine Chapters* require taking square roots using the square root algorithm. To take the square root of a $2k + 1$ or $2k + 2$ digit number N , the algorithm begins by finding the largest number $A_0 = a_0 \times 10^k$, where a_0 is a digit, such that $A_0^2 \leq N$. Then compute $N_1 = N - A_0^2$. Now find the largest $A_1 = a_1 \times 10^{k-1}$ such that $A_1(2A_0 + A_1) \leq N_1$, and form $N_2 = N_1 - A_1(2A_0 + A_1)$. Continue in this manner. If N is a perfect square, its square root will be the $(k + 1)$ -digit number $S = a_0 a_1 \cdots a_k$.

Liu Hui first gives a geometric argument, similar to arguments used in Greek geometric algebra, to explain why the algorithm works. Consider Figure 2, which is not to scale. (Liu’s original figures were all lost, but most of them are easy to reconstruct from his verbal descriptions.) From a square of area N , we first subtract a square of side A_0 , then the L-shaped figure of width A_1 , which the Greeks called a gnomon, then a gnomon of width A_2 , and so on until we exhaust the square.

Well, at least we exhaust the square if N is a perfect square, as it is in many of the *Nine Chapters* problems. (Some of the problems involve rational perfect squares, for instance $N = 564752\frac{1}{4}$ in problem 4.15.) But Liu also asks what happens if N is not a perfect square: “In this case it is not sufficient

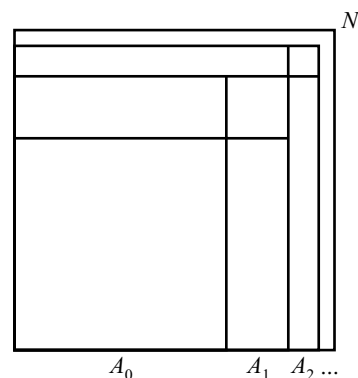


Figure 2. Geometry of the square root algorithm

to say what the square root is about by simply ignoring the [remaining] gnomon.”[7, p. 211] For integral but non-square N , the square root algorithm yields $N = S^2 + R$, where $0 < R < 2S + 1$. Liu gives two ways of approximating the square root. The first is to take a rational approximation using

$$S + \frac{R}{2S + 1} < \sqrt{N} < S + \frac{R}{S}. \quad [17]$$

The second is even more interesting. If we continue the algorithm on the counting board past the last digit of N , we get

$$\sqrt{n} \approx a_0 a_1 \dots a_k + \frac{a_{k+1}}{10} + \frac{a_{k+2}}{100} + \dots$$

The ancient Chinese had names for the fractions $1/10^k$ for k up to five. Liu suggests continuing the calculation down to “those small numbers for which the units do not have a name,” and if necessary adding a fraction to a_{k+5} to get even greater accuracy [11]. In other words, it is not stretching very much to say that Liu Hui invented decimals; he certainly invented their calculational equivalent. We will see that he needed this kind of accuracy for his calculation of π . Liu also gave a justification for the cube root algorithm using a three-dimensional figure similar to Figure 2.

Chapter Eight of the *Nine Chapters* solved systems of linear equations using the *fangcheng* method on a counting board matrix: multiples of rows (actually columns, since the equations were set up vertically on the counting board) were systematically subtracted from other rows to reduce the matrix to triangular form. Liu Hui explains that the goal of this method is to reduce to a minimum the number of computations needed to find the solution: “generally, the more economic a method is, the better it is.” In fact, Liu compares two different *fangcheng* methods for solving problem 8.18 by counting the number of counting board operations needed in each method [17]. Surely this is the first example in history of an operation count to compare the computational efficiency of two algorithms.

Finally, Chapter Nine of the *Nine Chapters* presented, without justification, solutions to a number of problems involving right triangles. Liu Hui justified these solutions by a series of ingenious “dissection” arguments, based on the principles that congruent figures have the same area, and that if we dissect a figure into a finite number of pieces, its area is the sum of the areas of the pieces. I’ll give two examples.

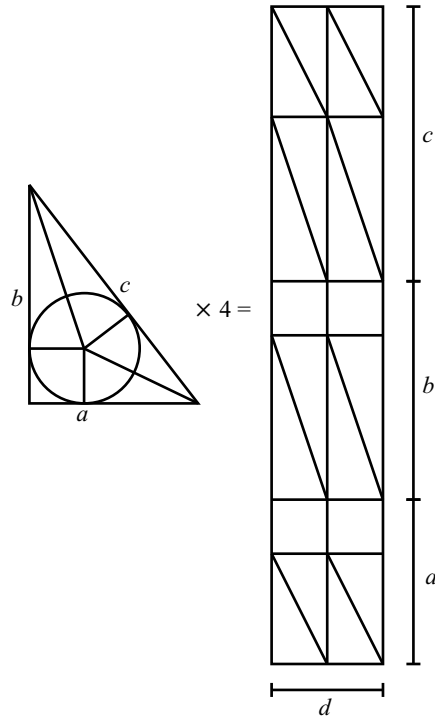


Figure 3. Diameter of a circle inscribed in a right triangle

The solution to problem 9.16 finds the diameter d of a circle inscribed in a right triangle with legs a and b and hypotenuse c by

$$d = \frac{2ab}{a + b + c}.$$

Liu’s dissection proof of this result can be reconstructed as in Figure 3 [20]. See it?

For the second example, consider the famous *gou-gu* theorem that for a right triangle as above, $a^2 + b^2 = c^2$. For this theorem, Liu’s verbal description of his proof is

The shorter leg multiplied by itself is the red square, and the longer leg multiplied by itself is the blue square. Let them be moved about so as to patch each other, each according to its type. Because the differences are completed, there is no instability. They form together the area of the square on the hypotenuse. [31, p. 71]

Clearly Liu had a dissection proof of the *gou-gu* theorem. Just as clearly, the verbal description does not enable us to reconstruct Liu’s diagram. Figure 4 shows two proposed constructions. The first, where the square on the hypotenuse is allowed to overlap the squares on the legs, is due to Gu Guangang in

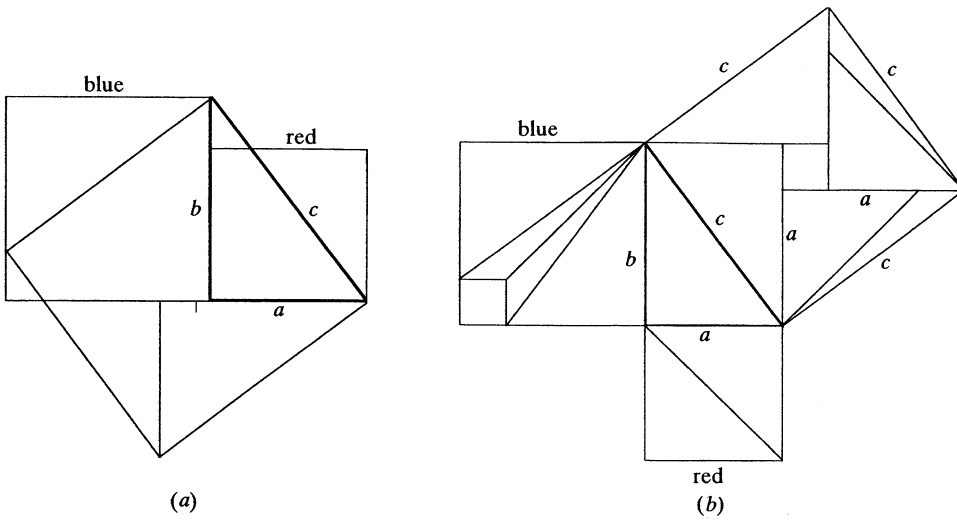


Figure 4. Dissection proofs of the *gou-gu* theorem

1892, reported in [17]. The second, less straightforward but without overlapping squares, is from [31].

5 The Sea Island Mathematical Manual

Chapter Nine of the *Nine Chapters* included surveying problems involving one unknown distance or length. However, most real surveying problems involve several such unknowns. For example, we might wish to determine the height of, and distance to, a mountain which is inaccessible, perhaps because it is on an island we cannot reach. Liu Hui pointed out that we can do this by making two observations, and worked out the geometry of how to make two observations yield the unknown distances. If we wish also to know the height of a pine tree on top of that inaccessible mountain, we can do it with three observations. His compilation of solutions to nine illustrative surveying problems became the *Sea Island Mathematical Manual*. The mountain on the sea island is the first problem; the pine tree is the second. [1] and [24] include complete translations with commentary.

Here is the sea island problem:

For looking at a sea island, erect two poles of the same height, 30 *chi*, the distance between the front and rear pole being 6000 *chi*. Assume that the rear pole is aligned with the front pole. Move away 738 *chi* from the front pole and observe the peak of the island from ground level;

it is seen that the tip of the front pole coincides with the peak. Move backward 762 *chi* from the rear pole and observe the peak from ground level again; the tip of the rear pole also coincides with the peak. What is the height of the island and how far is it from the front pole?

Answer: The height of the island is 7530 *chi*. It is 184500 *chi* from the front pole. [24, p. 20]

The extant version of the *Sea Island Manual* contains only the problems, answers, and recipes for obtaining the answers, exactly as in the *Nine Chapters*. Liu Hui also gave proofs for the correctness of his methods, but these proofs and the accompanying diagrams were not preserved, and the best we can do is offer plausible reconstructions. Using the notation of Figure 5, Liu’s method for solution corresponds to the formulas

$$h = x + b = \frac{bd}{a_1 - a_2} + b, \quad y = \frac{a_2d}{a_1 - a_2}.$$

We must obtain these formulas using only similar

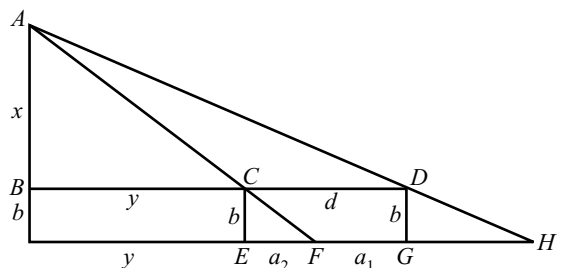


Figure 5. The height of a sea island

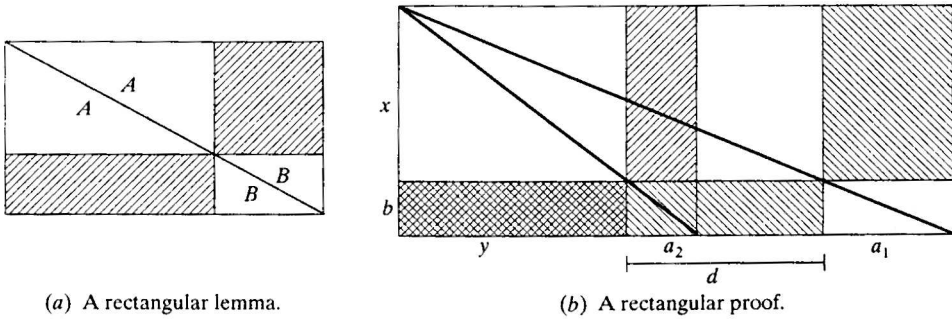


Figure 6.

right triangles, since there was no concept of angle, much less any trigonometry, in ancient Chinese mathematics, nor was there any use of similar triangles other than right triangles. Here is one method. Since $\triangle ABD \sim \triangle DGH$,

$$\frac{x}{y+d} = \frac{b}{a_1}, \text{ so } xa_1 = by + bd. \quad (1)$$

Since $\triangle ABC \sim \triangle CEF$,

$$\frac{x}{y} = \frac{b}{a_2}, \text{ so } xa_2 = by. \quad (2)$$

Subtracting these equations gives $x(a_1 - a_2) = bd$ which leads to the expression for the height, and then substitution gives the distance.

Swetz [24] gives a very plausible alternate derivation which avoids the use of similar triangles completely. It is based on a lemma about rectangles which is illustrated in Figure 6a: if we divide a rectangle into four smaller rectangles at any point on its diagonal, then the two rectangles shaded in the figure must have the same area. This follows from a dissection argument. The diagonal divides the rectangle into two congruent triangles. From these triangles, subtracting the congruent triangles labeled *A* and *B* yields the given rectangles. If we apply this result twice to Figure 6b, the equal $\backslash\backslash\backslash$ rectangles give equation (1), and the equal $\text{//}\text{//}\text{//}$ rectangles give equation (2). This method is also discussed in [9].

The *Sea Island Manual* was certainly not the deepest mathematics which Liu Hui did, but it probably had the greatest immediate impact. Recall that the kingdom of Wei was continually at war during the time of Liu’s work. Surveying was important for maps which supported war, as well as the administrative bureaucracy. Needham reports that the Wei general Deng Ai always “estimated the heights and distances, measuring by finger breadths before

drawing a plan of the place and fixing the position of his camp.” [24, p. 15] There is an interesting parallel in the West. Swetz notes that Greek armies had a specific reason for wanting to calculate unknown height at an inaccessible distance, quoting Heron of Alexandria:

How many times in the attack of a stronghold have we arrived at the foot of the ramparts and found that we made our ladders and other necessary implements for the assault too short, and have consequently been defeated simply for not knowing how to use the Dioptera for measuring the heights of walls; such heights have to be measured out of the range of enemy missiles. [24, p. 28]

6 The calculation of π

Recall that problem 1.32 of the *Nine Chapters* gave the correct formula for the area of a circle, but used a value of three for π . Liu points out that for a circle of radius one, the area of a regular dodecagon inscribed in the circle is three, so the area of the circle must be greater than three. He then proceeds to estimate the area of the circle more exactly by calculating the areas of inscribed $3 \cdot 2^n$ -gons as follows. In a circle of radius r , let c_n be the length of the side of an inscribed n -gon, a_n be the length of the perpendicular from the center of the circle to the side of of the n -gon, and S_n be the area of the n -gon. See Figure 7. Then we can calculate inductively

$$\begin{aligned} c_6 &= r, \\ a_n &= \sqrt{r^2 - (c_n/2)^2}, \\ c_{2n} &= \sqrt{(c_n/2)^2 + (r - a_n)^2}, \\ S_{2n} &= \frac{1}{2}nr c_n. \end{aligned}$$

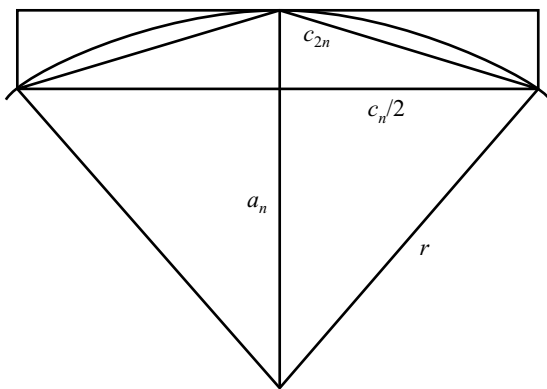


Figure 7. The calculation of π

The last formula is clever, and follows from noticing that each of the $2n$ triangles making up the $2n$ -gon can be thought of as having base r and height $c_n/2$. Moreover, Figure 7 shows that the area S of the circle satisfies

$$S_{2n} < S < S_n + 2(S_{2n} - S_n) = 2S_{2n} - S_n.$$

Liu considers what happens when we take n larger and larger: “the finer one cuts, the smaller the left-over; cut after cut until no more cut is possible; then it coincides with the circle and there is no left-over.” [20, p. 347] As n gets large, S_{2n} approaches the area of the circle and nc_n approaches the circumference, so we have justified the *Nine Chapters* claim that the area of a circle is one-half the product of its radius and circumference.

Taking $r = 10$, Liu Hui carries out the calculations, keeping 6-place accuracy, up to $n = 96$, hence approximating the circle by a 192-gon. He concludes that

$$3.1410 < \pi < 3.1427,$$

and suggests that for practical calculations it should be enough to use $\pi \approx 3.14$. Either Liu or some interpolating later commentator carried the computation as far as $n = 1536$ and obtained the approximation $\pi = 3.1416$. See [13] and [28] for treatments of the intricacies of this kind of calculation. [13] gives a translation of Liu Hui’s text.

If we compare this treatment to Archimedes’ in *Measurement of a Circle*, the similarities are striking, although the differences are also interesting. Archimedes, of course, included a formal proof by the method of exhaustion required by the conventions of Greek geometry. However, the subdivision method and the inductive calculation are essentially the same. Archimedes obtained his upper bound

by considering circumscribed polygons, instead of Liu’s clever method of using only inscribed polygons. Archimedes used 96-gons to obtain his famous estimate

$$3\frac{10}{71} < \pi < 3\frac{1}{7}, \text{ or } 3.1409 < \pi < 3.1428.$$

Two centuries later Zu Chongzhi (429–500 A.D.) carried Liu Hui’s approach farther. Using a polygon of 24576 sides, Zu obtained the bounds $3.1415926 < \pi < 3.1415927$. See [13] and, for a different view, [28]. In addition, Zu recommended two rational approximations for π , Archimedes’ value of $22/7$, and the remarkably accurate $355/113 \approx 3.1415929$.

Zu’s method for arriving at his rational approximation $\frac{355}{113}$ for π is not known. One line of reasoning would be to start with Zu’s value of 3.1415926 and the approximation $\frac{22}{7} = 3\frac{1}{7} \approx 3.1428571$, which is slightly too large, and ask for a fraction which, when added to 3, would give a better approximation than $\frac{1}{7}$ does. It is easy to see that the fractions we should check are those of the form $\frac{k}{7k+1}$. We then try to find k so that

$$\begin{aligned} \frac{1}{7} - \frac{k}{7k+1} &\approx .1428571 - .1415926 = .0012645, \\ \frac{1}{49k+7} &\approx .0012645, \quad 49k+7 \approx 791. \end{aligned}$$

The solution $k = 16$ gives the rational approximation $3\frac{16}{113} = \frac{355}{113}$. For another possible approach, see [17].

Zu Chongzhi’s approximation of π was not bettered until al-Kashi of Samarkand computed π to 14 decimal places in the early 15th century. The rational approximation $355/113$ was not discovered in Europe until the late 16th century.

7 The volume of pyramids

Chapter Five of the *Nine Chapters* gives correct formulas for the volumes of a number of pyramidal solids. For example, the volume of the *chu-tung*, a truncated rectangular pyramid illustrated in Figure 11, is correctly given as

$$\frac{h}{6}(2ab + ad + bc + 2cd).$$

Did you know that formula? From it follows the volume of a rectangular pyramid (put $c = d = 0$), a truncated square pyramid (put $a = b, c = d$), and a rectangular wedge (put $d = 0$).

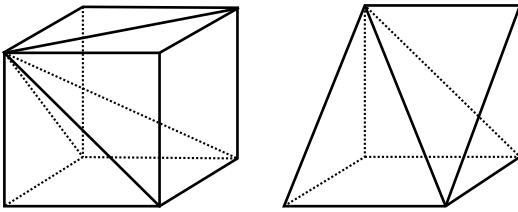


Figure 8. Dissecting a cube and a *qiandu*

Liu Hui gives justifications for these formulas based on dissection arguments and a remarkable limit argument. I will mostly follow the translation and discussion in [30]. Liu’s argument uses three special solids: a *qiandu*, which is a triangular prism, a *yangma*, which is a rectangular pyramid whose vertex is above one corner of its base, and a *bienao*, which is a tetrahedron with three successive perpendicular edges. See Figures 8, 9 and 10.

Liu starts with the case of a cube, which he dissects into three congruent *yangma*, to conclude that the volume of a regular *yangma* is $1/3$ the volume of the cube. See Figure 8. Since a *yangma* and a *bienao* fit together to make a *qiandu*, which is $1/2$ of the cube, the volume of the *bienao* must be $1/6$ the volume of the cube. Alternatively, we could get this result by dissecting the *yangma* into two congruent *bienao*.

Now suppose that instead of a cube, we start with an $a \times b \times c$ rectangular box. We can still dissect it

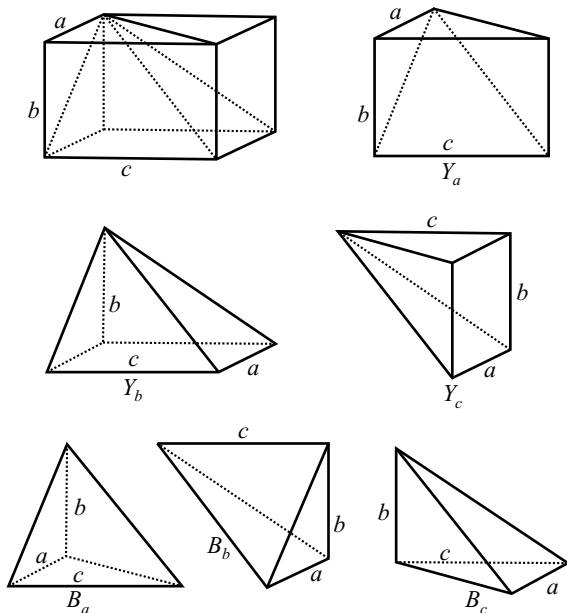


Figure 9. Three types of *yangma* and *bienai*

into three *yangma*, but now these *yangma* will have 3 different shapes, so it is not clear that their volumes are equal. We can also dissect a *yangma* into two *bienao*, or assemble a *bienao* and a *yangma* to make a *qiandu*, but again, the *bienao* have 3 different shapes, and it is not clear that their volumes are equal. Using the notation in Figure 9, what the dissections do show is that

$$Y_a + Y_b + Y_c = abc$$

$$\begin{aligned} Y_a + B_a &= abc/2 & Y_a &= B_b + B_c \\ Y_b + B_b &= abc/2 & Y_b &= B_a + B_c \\ Y_c + B_c &= abc/2 & Y_c &= B_a + B_b. \end{aligned}$$

However, this does not give enough information to evaluate the volumes.

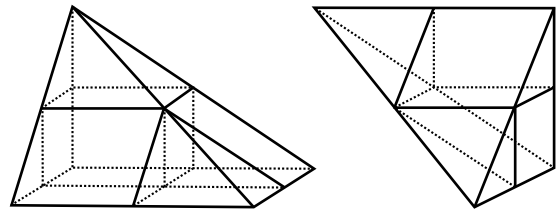


Figure 10. Dissecting a *yangma* and a *bienao*

Liu proceeds to prove that $Y_b = 2B_b$ (and similarly $Y_a = 2B_a, Y_c = 2B_c$), which does allow us to conclude that the volume of each *yangma* is $abc/3$ and that of each *bienao* is $abc/6$. His method is shown in Figure 10. Dissect Y_b at the midpoints of its sides into a rectangular box, 2 *qiandu*, and two half-size copies of Y_b (call them Y'_b). Similarly, dissect B_b into 2 *qiandu* and 2 half-size copies of B_b (call them B'_b). Since the box and 2 *qiandu* have twice the volume of 2 *qiandu*, we only need to show that $Y'_b = 2B'_b$. Liu notes that these new figures together have $1/4$ the volume of the original figures, since the two small *yangma* and *bienao* fit together to form two *qiandu* whose total volume is $abc/8$. Repeat the dissection on each of the new figures, and continue. At each stage the volume we have not yet accounted for is $1/4$ that of the previous stage. Liu expresses what happens in the limit as follows:

The smaller they are halved, the finer are the remaining dimensions. The extreme of fineness is called minute. That which is minute is without form. When it is explained in this way, why concern oneself with the remainder? [30, p. 173]

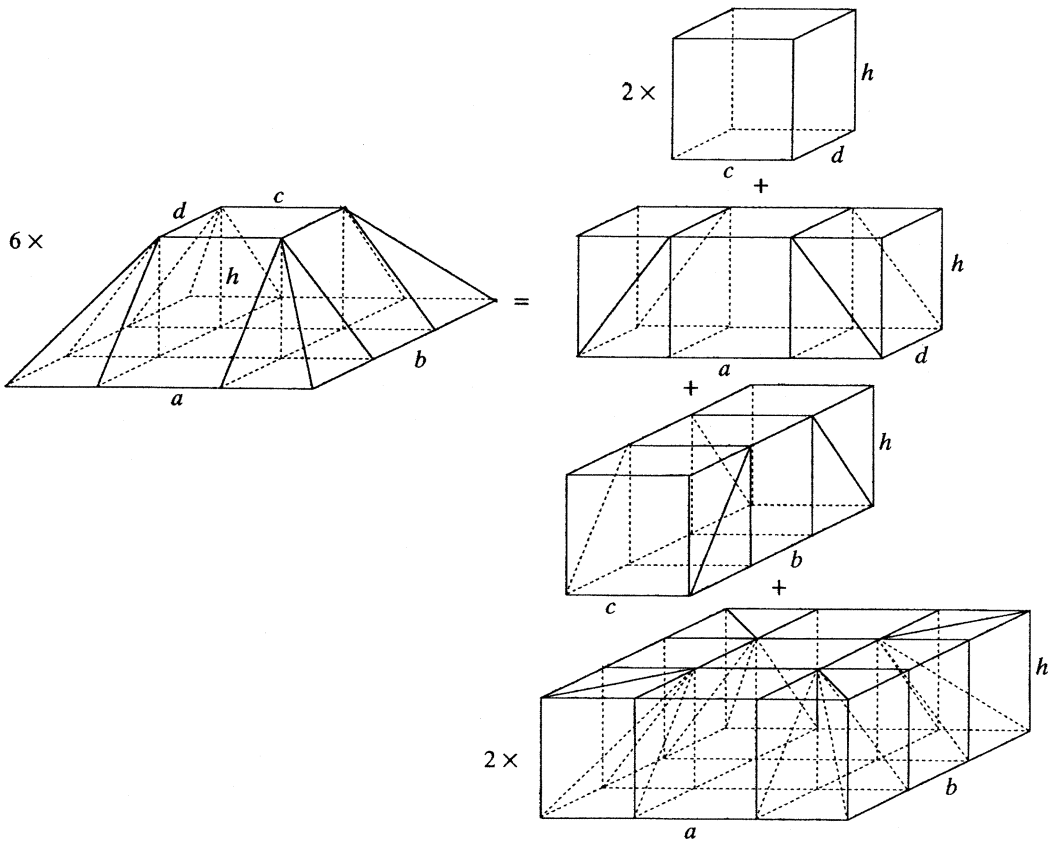


Figure 11. The volume of a *chu-tung*

This is not a modern limit argument, of course. Liu seems to be saying that if we cut the figures into smaller and smaller pieces, we will come to a point where the pieces are so small that they no longer have form or volume. (The terms translated as ‘minute’ and ‘form’ are philosophical terms from the *Tao Te Ching*.) Still, we recognize the limit idea, and the recursive dissection argument has a delightful elegance. For some of the philosophical issues, see [7], [16] and [30]. For a comparison to the Greek proof in Euclid’s *Elements*, see [4].

Knowing the volume of a *yangma*, we can now derive the volumes of the other solids by dissection. For example, let’s verify the formula for the volume of the *chu-tung*. Dissect it as in Figure 11 into a box L , four *qiandu* of two different shapes Q_a and Q_b , and four *yangma* Y . If we do this to six copies of the *chu-tung*, we have

$$6L + 12Q_a + 12Q_b + 24Y.$$

Now reassemble these, as in Figure 12, into two boxes of volume hcd : $2L$

- one box of volume had : $L + 4Q_b$
- one box of volume hbc : $L + 4Q_a$
- two boxes of volume hab : $2L + 8Q_a + 8Q_b + 24Y$.

Notice that for the last step we need to replace some of the Y_h *yangma* with *yangma* of other shapes, but this is allowable since we have shown that these *yangma* all have the same volume.

Finally, Liu derives the volume of a cone from the volume of a square pyramid, and the volume of a truncated cone from the volume of a truncated square pyramid, by using what we know as “Cavalieri’s principle,” after Bonaventura Cavalieri (1598–1647). We can state this principle as

The volumes of two solids of the same height are equal if their planar cross-sections at equal heights always have equal areas; if the areas of the planar cross-sections at equal heights always have the same ratio, then the volumes of the solids also have this ratio.

Liu inscribes the truncated cone, for example, in a truncated square pyramid of the same height, and

then says that since each cross-section consists of a circle inscribed in a square, the ratio of the volumes of the truncated cone to the truncated pyramid must be in the same ratio as the area of a circle to its circumscribed square, i.e., $\pi/4$ [7].

8 The volume of a sphere

Recall that problem 4.24 of the *Nine Chapters* gave the volume of a sphere as $\frac{9}{16}d^3$. Liu points out that this is incorrect, even using the inaccurate value of 3 for π . He explains the error as follows. Let a cylinder be inscribed in a cube of side d , and consider the cross-section of this figure by any plane perpendicular to the axis of the cylinder. The plane will cut the cylinder in a circle of diameter d , inscribed in a square of side d . The ratio of these areas is $\pi/4$. Since this is true for each cross-section, the same ratio must hold for the volumes, so that the volume of the cylinder is $\frac{\pi}{4}d^3$. Now consider the sphere of diameter d inscribed in the cylinder. If we assume, incorrectly, that the ratio of the volume of the sphere to the volume of the cylinder is also $\pi/4$, then we get that the volume of the sphere is $\frac{\pi^2}{16}d^3$, which is the *Nine Chapters* result (using $\pi = 3$).

How do we know that the ratio of the volumes of the sphere and cylinder cannot be $\pi/4$? Liu's ingenious argument is as follows. Inscribe a second cylinder in the cube, with axis orthogonal to that of the first cylinder, and consider the intersection of these two cylinders. Liu called this intersection a "double box-lid." See Figure 12. Since the sphere is contained in both cylinders, it is contained in the box-lid. Moreover, consider any cross-section of this

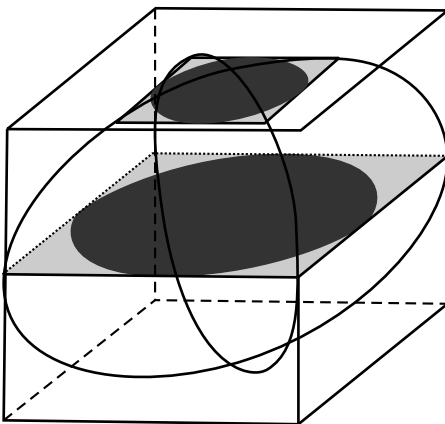


Figure 12. Cross sections of a sphere in a double box-lid in a cube

figure by a plane perpendicular to the axis of the box-lid. The cross-section of the sphere will be a circle, inscribed in the square which is the cross-section of the box-lid, so again the ratio of the areas is $\pi/4$, and since this is true for all cross-sections, the ratio of the volumes of the sphere and the box-lid must also be $\pi/4$. Now the box-lid is certainly smaller than the original cylinder, so the ratio of the volumes of the sphere and the cylinder must be strictly less than $\pi/4$.

This lovely argument using Cavalieri's principle shows that the *Nine Chapters* formula is wrong, but in order to use it to find the correct volume of the sphere, we would need to be able to find the volume of the double box-lid. Liu tried to do this, but could not. He recorded his failure in a poem, translated by D. B. Wagner as "The Geometer's Frustration:"

Look inside the cube
And outside the box-lid;
Though the diminution increases,
It doesn't quite fit.
The marriage preparations are complete;
But square and circle wrangle,
Thick and thin make treacherous plots,
They are incompatible.
I wish to give my humble reflections,
But fear that I will miss the correct principle;
I dare to let the doubtful points stand,
Waiting for one who can expound them.

[29, p. 72]

The wait turned out to be two centuries, and the person Liu waited for was Zu Gengzhi, the son of Zu Chongzhi. Stories associated with Zu Gengzhi are reminiscent of those told about Archimedes and many mathematicians since then. For instance, "he studied so hard when he was still very young that he did not even notice when it thundered; when he was thinking about problems while walking he bumped into people." [15, p. 82]

Zu Gengzhi argues as follows. Consider one eighth of the double box-lid inscribed in the cube of side $r = d/2$. See Figure 13. If a plane is passed through this figure at height h , it intersects the cube in a square of side r , and the box-lid in a square of side s . By the *gou-gu* theorem, $r^2 - s^2 = h^2$. Hence the area of the gnomon *outside* the box-lid is h^2 .

Now Zu Gengzhi considers another solid of height r whose cross-section at height h is h^2 : an inverted *yangma* cut from a cube of side r . See Figure 13. The part of the cube outside the box-lid, and this *yangma*, have all their corresponding cross-sections

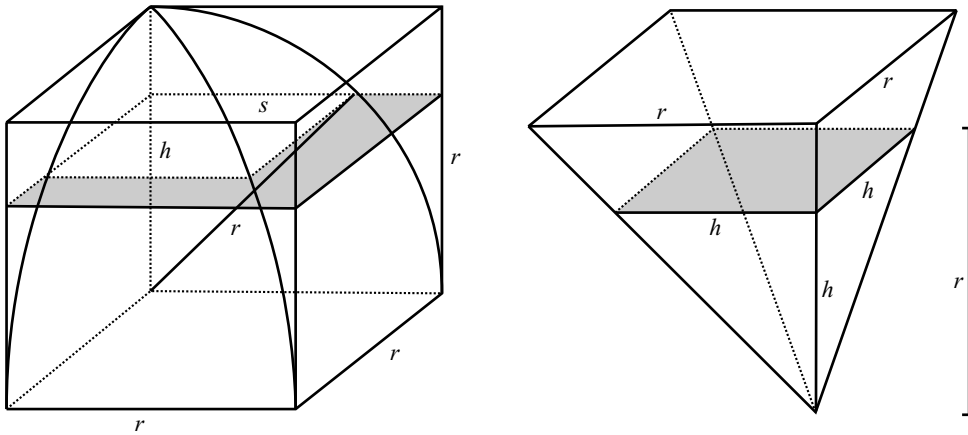


Figure 13. The volume outside a box-lid is Cavalieri-equivalent to a *yangma*

of the same area. Zu then states his version of Cavalieri's principle in verse:

If volumes are constructed of piled up blocks [areas],
And corresponding areas are equal,
Then the volumes cannot be unequal. [29, p. 75]

Since the volume of the *yangma* is $\frac{1}{3}r^3$, and the volume outside the box-lid must be the same, the volume inside the box-lid must be $\frac{2}{3}r^3$. Putting the eight pieces together, we get that the volume of the complete double box-lid must be two-thirds of the cube containing it, $\frac{2}{3}d^3$. Remembering Liu Hui's result that the sphere takes up $\pi/4$ of the double box-lid, we finally get the correct formula for the volume of a sphere of diameter d :

$$V = \frac{\pi}{4} \frac{2}{3} d^3 = \frac{\pi}{6} d^3.$$

Following Liu, Zu ends his discussion with a poem, "The Geometer's Triumph:"

The proportions are extremely precise,
And my heart shines.
Chang Heng copied the ancient,
Smiling on posterity;
Liu Hui followed the ancient,
Having no time to revise it.
Now what is so difficult about it?
One need only think. [29, pp. 76–77]

One could argue that Liu Hui did not use the full power of Cavalieri's principle, since he only applied it to the situation of one figure inside another, where the cross-sections were circles inscribed in squares. But certainly Zu Gengzhi gave a clear statement of the principle and used its power more than a millennium before Cavalieri [14].

There was another precursor, of course. Archimedes had calculated the volume of a sphere, and in Proposition 15 of *The Method*, he calculated the volume of the perpendicular intersection of two cylinders of the same radius. The argument for Proposition 15 is in the part of *The Method* which has not survived, but it is not difficult to reconstruct the reasoning from other demonstrations earlier in the book. Archimedes thought of volumes as made up of planar slices and balanced them on a lever against the slices of other volumes. It is an extension of Cavalieri's principle. For a general discussion of the use of versions of Cavalieri's principle in Greek geometry, see [10].

9 Conclusion

After the theoretical phase of Chinese mathematics in the 3rd through 5th centuries, represented by Liu Hui, Zu Chongzhi and Zu Gengzhi, proofs and justifications began to be less important. Although the work of Liu Hui was still taught in the official School for the Sons of the State, instruction began to emphasize rote learning of methods rather than justifications. Liu's diagrams from the commentary on the *Nine Chapters* and arguments from the *Sea Island Manual*, and Zu Chongzhi's work, were lost. The next, brief flowering of creative mathematics in China did not happen until the 13th century, with mathematicians like Qin Jiushao, Li Zhi, Zhu Shijie and Yang Hui. After the thirteenth century, Chinese mathematics declined again until the period of contact with the West.

It is interesting to speculate why Chinese mathematics, with such a powerful calculational base

and such a strong theoretical start, did not develop a coherent, ongoing mathematical tradition. Martzloff [17] and Swetz [25] review a number of possible reasons: emphasis on practical applications, rote learning and reverence for established ideas which stifled creativity, uneven state support, and low social status accorded to mathematicians compared to scholars in the humanities.

Nevertheless, the remarkable achievements of Chinese mathematics in its first golden age are worthy of our interest and admiration.

References

[8] and [21]–[27] contain very accessible introductions to Chinese mathematics. [15] and [17] are comprehensive modern histories of Chinese mathematics which make extensive use of Chinese research. [18] and [19] are older histories which are still good reading.

1. Ang Tian Se and Frank Swetz, A Chinese mathematical classic of the third century: The Sea Island Mathematical Manual of Liu Hui, *Historia Mathematica*, 13 (1986) 99–117.
2. David Burton, *History of Mathematics: An Introduction*, Wm. C. Brown, Dubuque, 1995.
3. Karine Chemla, Theoretical aspects of the Chinese algorithmic tradition (first to third century), *Historia Scientiarum*, 42 (1991) 75–98.
4. J. N. Crossley and A. W. C. Lun, The logic of Liu Hui and Euclid as exemplified in their proofs of the volume of a pyramid, *Philosophy and the History of Science: A Taiwanese Journal*, 3 (1994) 11–27.
5. Heath, T. L., ed., *The Works of Archimedes*, Dover, 1912.
6. Ho Peng-Yoke, Liu Hui, *Biographical Dictionary of Mathematicians*, Charles Scribner's Sons, New York, 1991.
7. Horng Wann Sheng, How did Liu Hui perceive the concept of infinity: a revisit, *Historia Scientiarum*, 4 (1995) 207–222.
8. G. Joseph, *The Crest of the Peacock*, Penguin, London, 1991.
9. Victor Katz, *A History of Mathematics: An Introduction*, Harper Collins, Chicago, 1993.
10. Wilbur Knorr, The method of indivisibles in ancient geometry, in R. Calinger, ed., *Vita Mathematica*, Mathematical Association of America, Washington, 1996.
11. Lam Lay Yong, Jiu Zhang Suanshu (Nine Chapters on the Mathematical Art): an overview, *Archive for History of Exact Sciences*, 47 (1994) 1–51.
12. Lam Lay Yong, Hindu-Arabic and traditional Chinese arithmetic, *Chinese Science*, 13 (1996) 35–54.
13. Lam Lay Yong and Ang Tian Se, Circle measurements in ancient China, *Historia Mathematica*, 13 (1986) 325–340.
14. Lam Lay Yong and Shen Kangsheng, The Chinese concept of Cavalieri's principle and its applications, *Historia Mathematica*, 12 (1985) 219–228.
15. Li Yan and Du Shiran, *Chinese Mathematics: A Concise History*, translated by J. Crossley and A. Lun, Oxford University Press, Oxford, 1987.
16. Geoffrey Lloyd, Finite and infinite in Greece and China, *Chinese Science*, 13 (1996) 11–34.
17. Jean-Claude Martzloff, *A History of Chinese Mathematics*, Springer-Verlag, New York, 1997.
18. Yoshio Mikami, *The Development of Mathematics in China and Japan*, Chelsea, New York, 1913.
19. Joseph Needham, *Science and Civilisation in China*, vol. 3: *Mathematics and the Sciences of the Heavens and the Earth*, Cambridge University Press, Cambridge, 1959.
20. Siu Man-Keung, Proof and pedagogy in ancient China: examples from Liu Hui's commentary on Jiu Zhang Suan Shu, *Educational Studies in Mathematics*, 24 (1993) 345–357.
21. Frank Swetz, The amazing Chiu Chang Suan Shu, *Mathematics Teacher*, 65 (1972) 425–430. Reprinted in F. Swetz, ed., *From Five Fingers to Infinity*, Open Court, Chicago, 1994.
22. —, The 'piling up of squares' in ancient China, *Mathematics Teacher*, 70 (1975) 72–79. Reprinted in F. Swetz, ed., *From Five Fingers to Infinity*, Open Court, Chicago, 1994.
23. —, The evolution of mathematics in ancient China, *Mathematics Magazine*, 52 (1979) 10–19. Reprinted in F. Swetz, ed., *From Five Fingers to Infinity*, Open Court, Chicago, 1994.
24. —, *The Sea Island Mathematical Manual: Surveying and Mathematics in Ancient China*, Pennsylvania State University Press, University Park, 1992.
25. —, Enigmas of Chinese mathematics, in R. Calinger, ed., *Vita Mathematica*, Mathematical Association of America, Washington, 1996.
26. Frank Swetz and T. I. Kao, *Was Pythagoras Chinese? An Examination of Right Triangle Theory in Ancient China*, Pennsylvania State University Press, University Park, 1977.
27. Robert Temple, *The Genius of China*, Simon and Schuster, New York, 1986.
28. Alexei Volkov, Calculation of π in ancient China: from Liu Hui to Zu Chongzhi, *Historia Scientiarum*, 4 (1994) 139–157.

29. D. B. Wagner, Liu Hui and Tsu Keng-chih on the volume of a sphere, *Chinese Science*, 3 (1978) 59–79.
30. —, An early Chinese evaluation of the volume of a pyramid: Liu Hui, third century A.D., *Historia Mathematica*, 6 (1979) 164–188.
31. —, A proof of the Pythagorean theorem by Liu Hui (third century A.D.), *Historia Mathematica*, 12 (1985) 71–73.

Number Systems of the North American Indians

W. C. EELLS

American Mathematical Monthly 20 (1913), 263–272, 293–299

The linguistic diversity of the Indians inhabiting the North American continent is one of the most remarkable features of world ethnology. The late director of the Bureau of American Ethnology says: “In philology, North America presents the richest field in the world, for here is found the greatest number of languages distributed among the greatest number of stocks.” [16, p. 78] The Bureau recognizes almost three score distinct linguistic families having no lexical resemblance, no apparent unity of origin, no relation to European or Asiatic languages. These “families” are further subdivided linguistically into 750 “tribes” or languages. [17, p. 1]

These languages differ as widely in *number words* and *number systems* as they do in other features. This is in marked contrast with the languages of the great Indo-European family where, even in languages which are mutually unintelligible, the same root words appear with great uniformity in the numerals. The very remarkable differences in the form and use of the numerals of the American Indians afford a fruitful field for study of the evolution of the concept of number among hundreds of distinct, uncivilized peoples. This paper is based upon an examination of the number systems of more than three hundred of these languages in North America. We will discuss the origin of number words and their principles of formation, the way in which they were built up into number systems, and some of the variations of these systems in actual use.

1 Principles of formation

1.1 Digital origin

The child’s most natural counters are his fingers; to them he turns almost instinctively when wishing

to count. What evidence is there that primitive peoples, races in the childhood state, have also turned to their digits for assistance? The answer to this question will throw much light on the origin of number words and their development into systems. We shall consider three kinds of evidence.

Evidence from systems used. The almost universal prevalence of decimal, quinary, or vigesimal systems of numeration on the North American continent is perhaps the strongest general evidence that counting in its origin is digital. But the octonary, quaternary and ternary systems mentioned later will show that such evidence is neither universal nor conclusive. To this indirect evidence, based on the systems used, can be added direct proof from observation and from the ascertained meaning of number words.

Observational evidence. Many observers report that Indians in various parts of the continent use their fingers, or fingers and toes, in counting, at the same time speaking the corresponding number words. With some tribes the *use* of the fingers is the important thing, the accompanying vocal utterance being of secondary importance; e.g., some of the Eskimo tribes use the same *words* for 6, 7, 8, 9, 10 as for 1, 2, 3, 4, 5 but count them on the second hand. In others the language development has led to independent numerals which often preserve evidence of their digital origin. Examples are given in the next section. In widely separated tribes all over the continent actual finger counting has been observed and the rather remarkable fact noted that the order of counting is almost always uniform, commencing with the little finger of one hand and counting to the thumb, thence to the thumb of the other hand and to the little finger again. Usually the fingers are bent

as the counting continues, but sometimes the hand is first clenched and the fingers then extended one at a time. [1, vol. 1, pp. 6–7] This general uniformity of order gives considerable aid in the linguistic discussion of the next paragraph.

Linguistic evidence. The German philologist Grimm in speaking of the Old World languages, says, “All number words come from the fingers of the hand.” [9, p. 167] Is this observation true for the New World? In the languages of civilized nations the numerals are so ground down from long usage that it is difficult to detect in their form their possible digital origin. While this is also the case in some Indian languages, in many there are striking similarities, while in others digital words and number words are almost or quite identical. Following the observed order of counting, the little fingers would be used for 1 and 10, the fourth fingers for 2 and 9, etc. Only a few typical examples of the many noted will be given with reference to each number.

ONE. Some of the names given the little finger are “the smallest,” “the last of the hand,” “little daughter of the hand.” While we do not expect to find the word for *one* always or even usually connected with that for little finger (since the concepts of unity doubtless preceded formal counting) yet some instances are known; e.g., Massachusetts: *pasuk* from *piasuk*, “very small”; Montagnais: *inlare*, “end is bent”; Zuni: *topinte*, “taken to start with.”

TWO is sometimes derived from finger; e.g., Montagnais: *nake*, “another bent in”; Dakota: *nonpa*, “to bend down”; Zuni: *kwillin*, “that (finger) put down with its like.” But more often it is connected with the word for “hand,” probably because there are two hands.

THREE. The third finger is often named “the middle”; e.g., Massachusetts: *nishwe*, from *nashaue*, “half way”; Zuni: *hain* “equally dividing one.”

FIVE is counted on the thumb and we have Karankawan: *natsa behema*, from *natsa*, “one,” *behema*, “finger.” But more prominent is the idea that the hand is completed, variously expressed as “finished,” “fingers finished,” “all fingers,” “all done,” etc.; e.g., Ojibwa: *nanan*, “gone,” “spent,” and similarly in several Eastern languages. Hidatsa: *kichu*, from *ki*, “completely,” *chu*, “turned down”; Ute: *munugi*, from *manoku*, “all.” In most instances however it is connected with “hand,” or “whole hand”; e.g., Kaniagmiut: *talgamen*, from *tallega*, “hand”; Comanche: *mowaka*, from *mowa*, “hand”;

Klamath: *tunep*, from *tu*, “away,” *nep*, “hand,” i.e., “hand-away.”

From six to nine the numerals are expressed (a) from the names of the fingers used; (b) by “hand” + 1, 2, 3, 4; (c) by 1, 2, 3, 4, + “again” or “besides” (most frequent); or by (d) by 1, 2, 3, 4 repeated without change. This last method is found only among the Eskimo and is probably always accompanied by the actual use of the second hand.

SIX. The Point Barrow six well illustrates the evolution of a number word. We are given the three forms *atautyimin-akbinigin-tudlimut*, literally, “once-on next- (and) five,” *atautyimin-akbinigin*, “once-on next,” *akbinigin*, “on next.” Other examples of six are Tano: *manli*, from *man*, “hand,” *li*, “piece,” i.e., “hand and piece of next”; Klamath: *nadshk-shapta*, “one I have bent over”; Takelma: *maimis*, “finger one in.”

SEVEN falls on the index finger or “pointer,” e.g., Zuni: *tserucek* from *tserverc*, “to point”; Greenland: *arfinek-mardluk*, “on the other hand-two”; Omaha: *penompa*, from *pe*, “finger,” *nompa*, “two.”

For **EIGHT**, Hudson’s Bay: *kittukleemot*, “middle finger”; Omaha: *pethatbathi*, “finger-three”; Klamath: *ndan-kshapta*, “three I have bent over.”

For **NINE** the subtractive principle comes into use and we have the additional forms “one left,” “only one,” etc. We also have the forms, Greenland: *mikkelerak*, “fourth finger”; Zuni: *tenalikya*, “all but one held up with rest.”

TEN is counted on the little finger, e.g., Hudson’s Bay: *eerkitkoka*, “little finger.” But more prominent is the fact “two hands completed,” “man finished,” or “man.” Thus Zuni: *astemthla*, “all of the fingers”; Wintun: *pampa-sempta* from *pampu-ta*, “two,” *sem*, “hand”; Konkau: *machoko*, from *mar*, “hand,” *choko*, “double.”

Above ten, various combinations of the first ten numerals occur in which of course these digital names reappear. A few other examples of interest will be given.

ELEVEN. Unalit: *atkhakhtok*, “it goes down” referring to change from hands to feet).

THIRTEEN. Greenland: *arkanenpingasut*, “on the first foot, three,” etc.

SIXTEEN. Unalit: *gukhtok*, “it goes over” (to toes of other foot).

NINETEEN. Maidu: *tsoi-ni-maiduk*, from *maidu*, “man,” *tsoi*, “four”—“four with man,” i.e., after 15, 4 on toward 20 (man); and similarly for 16, 17, 18.

TWENTY. In decimal-system languages twenty is usually but not always “two tens.” In the viges-

imal it is quite commonly “man,” “Indian,” “all hands and feet”; e.g., Navaho: *natin*, from *tine*, “man”; Greenland: *inuk-mavdlugo*, “man come to an end,” or *inup-avatai-navdlugit*, “man’s outer members completed”; Kaniagmiut: *swinuk*, from *suke*, *inuk*, “man”; Wintun: *ketet-wintun*, from *ketet*, “one,” *wintun*, “Indian”; Tuolomne: *rengé mewoom*, “one man”; Maidu: *kom maiduk*, from *maidu*, “Indian,” and *kom*, possibly “whole”; Shasta: *tsec*, from *tsec*, “man”; Tlingit: *tlekha*, from *tle*, “one,” *hka*, “man.”

Extent and distribution. Clear linguistic digital evidence similar to the examples given above has been found in about 40 per cent of the languages examined, uniformly distributed over the continent. Doubtless further study will reveal similar evidence in other languages especially where adequate vocabularies have not been available.

Non-digital evidence. We turn now to a consideration of the evidence that the origin of number words was non-digital in some languages. There are four phases to be considered.

(1) *First Four Numerals.* The concepts of unity and duality are so fundamental that in many instances we may be sure they were named before formal finger counting gave names to the corresponding words. *One* has a connection with the first personal pronoun in some languages. *Two* seems often to come from roots denoting separation, “that” as distinguished from “this,” or from ideas of pairs, being frequently related to the words for hands, feet, eyes, wings, husband and wife. *Three* is more frequently digital, but it seems sometimes to have a meaning of “more,” “many,” a plural as distinguished from a dual. Compare Micmac: *tchicht*, “three,” with the cognate Delaware *tchitch*, “still more.” *Four* is sometimes expressed by a word meaning “complete,” “right,” “perfect.” Its frequency as a sacred number among the North American Indians and its use in some cases as the base of a quaternary system indicate that it is a unique word of non-digital origin.

(2) *Arithmetical Operations.* Numbers higher than ten and in many cases those higher than five are expressed by arithmetical operations, and the digital meaning, even if present in the beginning, usually sinks into the background. The process of such combination begins earlier than the English in many Indian languages. We have numerous examples of $3 = 2 + 1$, $4 = 2 \times 2$, $4 = 2 + 2$, $6 = 3 \times 2$, $8 = 4 \times 2$, $10 = 5 \times 2$, $12 = 6 \times 2$, $9 = 3 \times 3$

and other rarer combinations. Thus there are many cases in which words for numerals above three are derived by purely arithmetical processes. Of course there are the higher numerals, hundred, thousand, million, where they exist, in which we should rarely expect digital evidence.

(3) *Marks of Completion.* When the Indian has counted ten or twenty he may use some reminder of the fact, such as a pebble, stick, arrow, grain of corn, etc. For example, Huchnom: 20, *pualya*, “one-stick-stand” and similarly for 40 and 60; in the same language 100 is *pual*, “one-stick” and similarly for 200; Maidu: 20, *penim nokom* “two arrow”; Gallinero: 100, *tcacuto-hai*, “ten-stick.”

(4) *Superlative and Indefinite.* When a simple arithmetical combination is not used, especially for the expression of higher units, a superlative principle is sometimes found. Hundred is often expressed, “big ten,” and thousand as “old hundred,” “big hundred” or “too many to count”; e.g., Delaware: 1,000, *ngutti kittapachkei*, “the great hundred”; Choctaw: 1000, *tahlepa siponki*, “old hundred”; Kwakiutl: 1,000,000 *tlinhi*, “number which cannot be counted”; compare the Greek “myriad.”

We conclude that in North American Indian languages it is by no means true that number words, even as far as ten, always “come from the fingers,” although they probably do in a large majority of cases and the close connection can be traced in many instances. There is little uniformity as to method of formation, considerable diversity being found even in adjacent languages of the same family. This would indicate that their separation into tribes preceded the development of formal counting.

1.2 Additive principle

Cantor says that addition and multiplication are two methods of counting as old as the formation of number words. [1, vol. 1, p. 8] The additive principle is found of course in all numeral systems. Three phases of it are of interest in the American Indian languages.

Repetition. This is the simplest form of the additive principle. If “one” is given, either as a symbol or as a word, “two” may be expressed “one-one,” “three” as “one-one-one,” etc., or by symbols as in the Roman numerals from one to four. In the gesture language of the Indians this is the method used, the fingers being the counters. In spoken languages no instance has been found of “two” as “one-one,” but there are several of “four” as “two-two”; e.g.,

Catawba: 2, *purra*, 4, *purrapurra*. In the Indian pictographs or hieroglyphics the simple repetition of strokes or notches is used, even for numbers up to a hundred. Sometimes these are grouped into tens by longer strokes or larger notches.

Addition in a base. English does not begin to use the additive principle until ten is reached, but many Indian languages begin much earlier. The earliest instance found is in the Coahuiltecan: 1, *pil*, 2, *ajtic*, 3, *ajtic-pil*. In other languages we find such expressions as “6-2 added” for 8, “8+1” for 9, “12+3” for 15, and of course very often “5+1, 2, 3, 4,” for 6, 7, 8, 9 and “10+1, 2, . . . , 9” as in English for 11, 12, . . . , 19. Those from 15 to 19 are also represented by “15+1, 2, 3, 4.” An interesting variation is shown by the Maidu numerals from 16 to 19 which in translation are “one with man” for 16, “two with man” for 17 and similarly for 18 and 19 to 20 “man,” the thought being “15 and one more on toward entire man.”

Precedence. Hankel and Fink call attention to a general law by which the written representation of numbers, when not confined to the mere rudiments, shows a tendency for higher numerals to precede the lower to represent addition. [10, p. 32; 6, p. 8] Is there a similar tendency for the spoken order of numerals among the Indian languages? Does the lower precede the higher or vice versa? In about 150 languages sufficient facts were available for study of the method of formation of compound numerals by the additive principle. The groups from 5 to 10, from 10 to 20, and above 20 have been considered separately. Using the notation “ $G + L$ ” to indicate “the greater is followed by the less” and “ $L + G$ ” for the opposite, “5+1,” “1+5,” etc., to stand for pure number combinations in the order given, and “1 + X ,” “ $X + 1$,” for the indefinite cases of an unknown element (probably non-numerical such as “again,” “besides,” etc.) combined with the known numbers, we may summarize the results of an examination of these languages as follows.

In the 5–10 group for the pure number combinations, $L + G$ and $G + L$ occur with about equal frequency. But if we include the unknown compounds $X + 1$ and $1 + X$ with $G + L$ and $L + G$ respectively, $L + G$ predominates about 2:1. In the 10–20 group for pure number combinations $G + L$ predominates strongly, 8:1, a reversal of the order of the first group. But for $X + 1$ and $1 + X$, $L + G$ predominates slightly, the ratio being 4:3, so that $G + L$ predominates in the group 2:1. In the group of higher combinations

$G + L$ predominates 16:1. Combining these results $G + L$ predominates altogether about 2:1. If the indefinite cases of $1 + X$ and $X + 1$ are excluded, only pure number combinations remaining, we approach close to a definite law. $G + L$ then predominates about 8:1. $G + L$ and $L + G$ both occur in the same language in fully half the cases examined. In this particular a marked contrast with the multiplicative principle will be observed. It may be mentioned that our own (oral) system is mixed, $L + G$ from 10 to 20, and $G + L$ for higher numbers.

1.3 Subtractive principle

Fink says: “In the verbal formation of a number system very rarely does subtraction come into use.” [6, p. 8] This statement is scarcely warranted for the systems of the American Indians for the principle has been found in 40 per cent of the languages examined. It occurs most frequently in expressions for “nine”; e.g., “one finger left,” “one from ten,” “one from finished,” etc., but also in the cases of 4, 7, 8, 14, 17, 18, 19 and the odd tens, 30, 50, etc. It is widely but not uniformly distributed over the continent. It occurs most frequently in the northern, eastern and central sections of the continent, less on the Pacific coast and least of all about the Gulf of Mexico. As is to be expected, “one subtracted” is most usual, occurring in about 30 per cent of the languages, “two subtracted” in about 5 per cent and “three subtracted” and “ten subtracted” in about 2 percent each. The use of the principle is found in some languages but not in closely related ones of the same family.

One subtracted. As examples we may give for *nine*, Unalit: *keka-mitatet*, “nearly ten,” or *payuk-ostau*, from *payuk*, “one,” *ostau*, “less?”; Uinta: *suromatampsuin*, “near-ten”; Haida: *klath-skwanon*, from *klath*, “one,” and *skwansin*, “ten.” For *four* we have Takhtam: *voatcham*, from *mah-at cham*, “five”; Zuni: *awiten*, “all fingers all but done.” For *fourteen*, Point Barrow: *akimiix-otaityuna*, “I have not quite 15.” For *nineteen*, Alaska: *enuenok-otalia*, “twenty less one.” *Thirty-nine*, Kulanapo: *pitikunamu-akhokaki*, “forty, one not.” Arikara expresses both 7 and 9 from 8 and 10 respectively by means of a diminutive particle, “little ten,” “little eight.”

Two subtracted. Onondaga: 8, *teg-ueron*, from *tegni*, “2”; Crow: 8, *nupa-pik*, from *nupa*, “2,” and *pirake*, “10”; Kwakiutl: *matl-gwanatl*, from *matl*, “2.”

Three subtracted. This occurs but rarely. Pawnee is noteworthy. In it, 17, 18, 19 are *tauit-kaki*, *pitkus-kaki*, *usku-kaki* from *usku*, "1," *pitkus*, "2" *tauit*, "3," and *kaki*, "less."

Ten subtracted. This is quite common among some of the California families, where the odd tens, 30, 50, 70, 90 are expressed as 40, 60, 80, 100 combined with ten, with the thought "60 lacking 10." In some cases however the thought may be "(40) and 10 on the way to 60," a variation of the additive principle not uncommon among these tribes.

1.4 Multiplicative principle

This principle, like the additive, is of universal use in Indian as well as in all other languages for the formation of higher numerals. It seems to begin in the Indian languages with the expression of 4 as "2×2." It is constantly used in the formation of "secondary bases" in the ternary, quaternary, quinary, decimal, and vigesimal systems. Our chief interest in studying this principle is in the question of precedence. A much more decisive law of precedence than in the case of the additive principle is found. About 200 languages afford data for the conclusion that the marked tendency in the formation of higher numerals by multiplication is for the lesser number to precede the greater. This coincides with both the oral and written English forms. In every group the type $L \times G$ predominates over the type $G \times L$ to a marked degree. For the formation of the tens, form of the type 2×10 predominate over the type 10×2 in the ratio 5:1; for the hundreds, 2:1; for the thousands, 2:1; for others (e.g., $8 = 2 \times 4$), 17:1. Or altogether a predominance of $L \times G$, 4:1. Almost all the instances of $G \times L$ are found in the languages about the Gulf of Mexico and in the single but large Siouan family of the plains. If the Siouan languages alone were left out of consideration the predominance of $L \times G$ would be about 8:1. In only five languages do we find $L \times G$ and $G \times L$ both occurring in the same system, a decided contrast to the results noted in the study of the additive principle.

1.5 Duplicative principle

A striking feature of the Indian numeral systems is the frequency of occurrence of a duplicative or pairing principle. In some instances 6 is expressed as "2×3," "again 3," "3,3," "threes" and similarly for 4, 8, 10, and even 12. The large number of natural pairs, such as the eyes, hands, arms, wings, etc., suggests that counting by pairs might be the course of

evolution followed by some languages. The standard histories of mathematics make little reference to such a principle as of any importance in the numeral systems of primitive people. [1, vol. 1, p. 5] We do find Hankel, however, making the rather surprising statement in discussing the number words of uncivilized people, that ten is never expressed as two times five, but always by a simple number word. [10, p. 20] This is not the case in American Indian languages; e.g., Gabrieleno: *wehes-mahar*, from *wehe*, "2," and *mahar*, "5"; Serranos: *wor-maharte*, from *maharte*, "5" *wor*, "2"; Patwin: *pampa-semta*, from *eti-semta*, "5," *pampet*, "2." Examples of its use for 4, 6, 8, 12 are Kutchin: 6, *neckh-kiethai*, 8, *nakhei-etanna*, from *nackhai*, "2," *kiethai*, "3," *etanna*, "4"; Kansas: 8, *kiya-tuba*, from *tuba*, "4," *kiya*, "again"; Shoshone: 4, *what-sowit*, from *what*, "2"; Cehiga: 12, *cape-nanba*, from *cape*, "6," *nanba*, "2"; Chwachamaju: 8, *kom-tca*, from *mitca*, "4," *ko*, "2."

In languages of the same family and even dialects of the same language there is variation in the use of this principle. And many examples can be given where the principle is not used at all in languages closely related to those in which it is. One feature is rather surprising; namely, of the approximately 125 cases of the probable use of this principle, it is far more common in the formation of four, six, eight than in the case of ten, even though ten is represented so commonly by the two hands. Fifty instances of its use for the formation of "8" are found, thirty-five for "4" (although some of these are somewhat obscure), twenty-five for "6," only ten for "10" and two for "12." Several languages seem to use it regularly in the formation of the even numbers to ten.

1.6 Divisive principle

Historians of mathematics agree on the rarity of the use of this principle in the formation of numerals the world over. [6, p. 8; 10, p. 21] Only two or three possible instances have been found in Indian languages. Thus we have Unalit: 10, *kolin*, and the literal meaning of the word, "upper half of the body"; Point Barrow: 10, *kodlin*, "upper part (of body)," and similarly in other Eskimo languages. One other example has been found, Pawnee: 5, *sihuks*, "hands half" from *ishu*, "hand," and *huks*, "half," i.e., "half of the two hands."

1.7 Fractions

Although many tribes had numeral systems of integers running into the thousands and even to millions,

very few of them had any idea of fractions. Where we do find such ideas they are of the most rudimentary sort. "One half" occurs most frequently, but only in about a dozen cases as far as noted, while examples of the other fractions are almost negligible. It is worth mentioning that the few instances we find are all of "unit fractions." Onondaga shows the best development of fractions, but how meager for a language whose numerals are given to one million. Its fractions are: $\frac{1}{2}$, *sat wachenonk*, meaning uncertain; $\frac{1}{3}$, *achen-na-degayagui*, from *achen*, "3," meaning "thrice divided"; $\frac{1}{4}$, *gayeri-degayagui*, from *gayeri*, "4," "four times divided."

1.8 Notation

As already mentioned, few symbols for number other than the spoken words are found. But on grave posts, buffalo robes, tattooing and other mnemonic pictographs there are a few pictured symbols. The usual method used for indicating numbers is by the repetition of single strokes, i.e., the additive principle in its simplest form. Sometimes the strokes are arranged in rows of ten or every tenth stroke is made longer than the others. Instances of this kind are found for the expression of numbers as high as thirty. They are found among the more highly civilized Indians of the middle west, especially among the Dakotas. On the other hand the Comanches are said to have been "ignorant of the elements of figures, even of a perpendicular stroke for one." [5, p. 416] Pure numeral notation is not always found. Frequently the number of objects is expressed by the repetition of the symbol for the object the desired number of times, especially in the case of men or tepees. Another method is by dots. A man's head with eight dots above it in one case means nine men, the head itself counting for one. Another picture gives a head over which are thirty black dots in three lines of ten each. This is said to mean thirty men, not thirty-one. Thus the usage is not fixed. The use of notches cut on sticks was frequent, not only in the middle west but in California and on Puget Sound in western Washington. At the San Gabriel mission in California every tenth notch was cut entirely across the stick instead of only in the corner.

2 Systems of numeration

Counting cannot be carried far by the use of successive unrelated terms of symbols for each number. For higher terms some system of compounding is

necessary, the usual method being by reference to some stopping point or *base* which is thought of as a new starting point. The choice of this base is of fundamental importance in the development of a true *number system* as distinguished from a mere *series of numbers*. The choice of this base seems usually to be related to primitive finger counting. One hand, two hands, or all the fingers and toes, are the three most natural stopping points, leading respectively to quinary, decimal, and vigesimal systems as illustrated below. There is much variation in the choice of these bases and in the way in which they are built up into systems for actual use after the base has been established. Thus a pure quinary system is rare, usually merging into a decimal or vigesimal one for the higher numbers. Accordingly in the list given below of the number of instances of each system found, quinary and quinary-decimal systems have been classed together, and similarly vigesimal and quinary-vigesimal. The numbers found are: decimal 146, quinary and quinary-decimal 106, vigesimal and quinary-vigesimal 35, quaternary 15, ternary 3, octonary 1 (binary 81). The binary instances given in parentheses refer simply to languages in which the duplicative principle (already discussed) occurs. These may be thought of as traces of a binary system. But there are no binary systems in the true sense of the word.

2.1 Decimal systems

Many of the American Indian systems have decimal scales as regular and as complete as our own, extending to quite high limits. Many others, while predominantly decimal, are combined with other systems, decimal-quinary and decimal-vigesimal being most frequent. The decimal system appears exclusively (i.e., no elements of other systems have been found) in only eleven families, all of which are small ones totaling only 19 languages of those examined. Thus in only a comparatively few cases do the languages of a single family use the decimal system consistently.

As illustrations we find "one hundred" expressed by a unique word or by such forms as "completed," "stock of tens," " 10×10 "; "thousand" by " $10 \times 10 \times 10$," " 10×100 ," "big hundred," "old man hundred," "large stock of tens"; "million" by " $1,000 \times 1,000$," "big thousand," "too many to count." While there are many variations, intermediate numbers as a rule are formed as in English.

2.2 Quinary and quinary-decimal systems

As already mentioned a pure quinary system is very rare, if indeed one exists. The pure form would require only five elements and is of course 1, 2, 3, 4, 5, $5 + 1$, $5 + 2$, $5 + 3$, $5 + 4$, 2×5 , and so on to 3×5 , 4×5 , and 5×5 as a new primary base. There are various approximations to such a system. Even 10 as " 2×5 " is uncommon although several instances of this have already been given. We have as variations for the numbers from 6 to 9, $6 = X + 1$ (" X " standing for some descriptive, non-numerical word), $7 = X + 2$, etc., arising from such forms of thought as "again 1," "second 1," "1 more," "on the other, 1," the numerals of the second quintate repeating without the use of the expressed base *five*. One or more of these forms may be entirely lacking, 8 being expressed " 2×4 " and similarly. But only in case at least two of the numbers between 5 and 10 show such a formation have we classed it as quinary. *Nine* is a frequent variation, since it is so often expressed by the subtractive principle. With these explanations we may state that quinary systems occur in about one third of the languages examined, appearing most frequently in the region around the Gulf of Mexico and least in the languages along the Northwest Coast.

Space will not permit illustrative systems to be given in full. Luiseno is purely quinary to ten, [4, p. 681] Gallinomoero has only a few variant forms until it reaches forty where it changes to decimal, [4, p. 676] while most of the quinary systems are of the general type, 1, 2, 3, 4, 5, $X + 1$, $X + 2$, $X + 3$, $X + 4$ and a new word for ten, e.g. Delaware [19, p. 65] or Shawnee. [13, p. 269]

2.3 Vigesimal and quinary-vigesimal systems

Vigesimal systems more or less complete occur in about one tenth of the languages examined. With the exception of the Caddoan family in the middle west they all appear along the Pacific Coast or in the far north. The obvious explanation of this system is in the digital origin of counting. The well known ethnologist Gatschet says that "tribes living in tropical and hot climates mostly possess the vigesimal system of the notation, which is rather infrequent among the Indians of the United States." [8, p. 210] He finds the explanation in the fact that they live barefoot as contrasted with the moccasined northern Indians. But this "barefoot" explanation, also given

by other writers (especially to account for the unusually well developed vigesimal systems of the Mexicans and Aztecs) [2], rather breaks down in the case of the Eskimo and tribes of the north Pacific Coast where the climate is scarcely adapted to the bare-foot stage. And yet among the Eskimo the vigesimal system has found its fullest development north of Mexico. The use of fingers and toes for the development of a vigesimal system seems to be independent of climate.

Twenty is the primary basis of the vigesimal system and is usually expressed as "man," "Indian," "man completed"; the multiples of twenty being expressed as "two men," "three Indians," etc. Pawnee carries this to 1,000 which is "50 persons." [3] The vigesimal system usually occurs in connection with some other system. Three types of combination may be noted.

Quinary-vigesimal. This is most frequent. The Greenland Eskimo says "other hand two" for 7, "first foot two" for 12, "other foot two" for 17 and similar combinations to 20, "man ended." The Unalit is also quinary to twenty which is "man completed." But 40 is "two sets of animals' paws," 60 "three sets of animals' paws" and so on regularly to 400 where there is an interesting change in the formation of this primary base (20×20) from animals back to man, for 400 is "20 sets of *man's* paws." [14, p. 238]

Decimal-vigesimal. Systems in which no quinary elements are found are comparatively rare. Wintun is alternately decimal and vigesimal, 20, 40, 60, being "one Indian," "two Indian," "three Indian," while 30, 50 are " 3×10 ," " 5×10 ." [4, p. 675] Others are similar.

Quinary-decimal-vigesimal. Several systems show a combination of the three digital bases in their formation. Kopiaimiut is quinary to ten, decimal for the formation of the odd tens and vigesimal for the even ones. [15] Amador is purely decimal to ten, quinary from ten to twenty, and then vigesimal, the odd tens being formed by addition of ten to the preceding even ones, e.g., $50 = 40 + 10$. [4, p. 680] Haida is quinary-decimal and quinary-vigesimal alternately to a hundred, then pure vigesimal, 400 being $20 \times 20 \times 1$ and 800 being $20 \times 20 \times 2$. [11, p. 123]

2.4 Quaternary systems

Fairly well-defined quaternary systems reaching to eight may be found among the Montagnais of the

far north, the Foxes of Wisconsin, the Iowas and Missouris of the Plains — but they find their best and fullest development into true systems in various California tribes. Usually they are mixed with other systems, but one or two cases are found of practically pure quaternary systems.

It is not easy to account for the origin of such a system. Two possibilities may be suggested.

(a) *Digital*. Perhaps a few tribes, for reasons best known to themselves, did not use their thumbs in counting. This is possible but there is little if any linguistic or observational evidence to support it. Most frequently these systems show the words, “stick,” “middle,” “body.” “Body” might be considered digital.

(b) *Sacred Number Theory*. Four was the sacred number of many widely separated tribes of Indians. The four cardinal points of the compass and the four seasons were recognized by the Indians. Among several tribes the literal meaning of “four” is “complete,” “right,” “perfect.” However the systems of these tribes are not quaternary ones. A chart for the sacred rites of the Ojibwas shows four degrees of initiation for medicine men. The swastika, with its four arms, originated with the Indians of the Southwest and was much used in basketry. There is a widespread “four worlds” of Indian mythology. In the various languages spoken on Puget Sound the same word for four, with minor variations, is seen most frequently, and it is the only number word common to about a dozen of these languages which have been most carefully studied by a missionary among them. All these data are suggestive of the origin of a quaternary system, although the fact remains that most of the instances given above are from tribes which do not actually possess such systems.

Santa Barbara as far as sixteen is as follows: 1, 2, 3, 4, $X+1$, $X+2$, $X+3$, 8, 9, $X+2$, 11, 3×4 , $X+1$, $X+2$, $X+3$, 16. For 20 and above it is decimal, probably due to contact with civilization. San Luis Obispo has $X+2$, $X+3$, for 6, 7; $8+1$ for 9; $12+1$, $12+2$, $12+3$ for 13, 14, 15. [4, p. 682] Numerous other languages have many quaternary forms.

2.5 Ternary system

A ternary system is much rarer than a quaternary and nowhere occurs in a pure form. Two well-developed ones are given below. It is even more difficult than in the case of the quaternary system to account for its origin. It is possible that to some primitive minds it seemed natural to count one, two, three,

and then by groups of threes. In some languages there is a similarity between the words for “this,” “that,” “that (remote)” or “here,” “there,” “yonder” and the first three numerals. A number of instances of 6 expressed as 2×3 have already been given and some writers have mentioned them as examples of a ternary system. But for reasons already explained they are better considered as formed by the Duplicative principle, unless occurring in connection with nine or twelve similarly formed. The Cuchan numerals for 3, 6, 9, are *ha-mook*, *hum-hook*, *hum-ha-mook*. [18, p. 41] In San Antonio 4 is related to 1, 6 is derived from 3, and 12 is directly 4×3 , 15 is 5×3 , and 13 is $12 + 1$. [4, pp. 683, 690]

But Coahuiltecan of Texas is the most interesting example found. It has binary, ternary, quaternary, quinary, decimal and vigesimal features, [7, p. 1] but seems to be prevalingly ternary. Its system as far as 50 is as follows:

1	4	$7 = 4 + 3$	$10 = 5 \times 2$
2	5	$8 = 4 \times 2$	$11 = 10 + 1$
$3 = 2 + 1$	$6 = 3 \times 2$	$9 = 4 + 5$	$12 = 4 \times 3$
$13 = 12 + 1$	$16 = 15 + 1$	$19 = 18 + 1$	$30 = 20 + 10$
$14 = 12 + 2$	$17 = 15 + 2$	20	$40 = 20 \times 2$
$15 = 5 \times 3$	$18 = 6 \times 3$		$50 = 40 + 10$

2.6 Octonary system

A single system with a base eight is known, the Yuki of California. This interesting system in translation is: [4, pp. 677, 685]

1	9 = beyond-1-hang	17 = 1-middle-project
2	10 = beyond-2-body	18 = 2-middle-project
3	11 = -3-body	19 = 3-middle-project
4	12 = -4-body	20 = 4-middle-project
5	13 = -5-body	
6	14 = -6-body	
7	15 = -7-body	
8	16 = middle-none	

2.7 Traces of other systems

Only traces of the use of other bases are found.

(a) *Binary*. If we include the large number of examples of the duplicative principle already discussed, we have many instances of binary elements. But these are only in the multiplicative principle and refer simply to doubling. The only place where we have found the additive principle used is in the Coahuiltecan (just given) where $3 = 2 + 1$. We may also notice Chutsinni: 2, *stunga*; 4, *stung-sung*; 8, *stun-sunga*. Yokaia: 2, *ko*; 4, *duo-ko*; 8, *ko-ko-dol*.

(b) *Sexanary*. Aside from the instances of twelve expressed as 2×6 mentioned under the duplicative principle, we have Wimunche Ute: $7 = 6 + 1$, Rumsen: $7 = X + 6$, $8 = 2 + 6$.

(c) *Base of Nine*. Trinity: $10 = 9 + 1$, $11 = 9 + 2$.

(d) *Base of Forty*. Chwachamaju: 40, *ku-hai*, “1-stick”; 80, *ko-hai*, “2-stick.”

(e) *Base of Sixty*. Achomawi: $70 = 60 + 10$, $80 = 60 + 20$.

2.8 Conclusions

The most striking feature of the systems which have been studied is their diversity, even in languages of the same family, and much more marked when the country as a whole is considered. For instance, in the closely related languages of the Yukian family in California, although the numerals from one to four are quite similar, yet two of the systems are quinary-decimal, a third is quinary-vigesimal, while the fourth is octonary; or in the Pujunan family in which one system is decimal, eight quinary-decimal and two quinary-vigesimal.

How many unique abstract number words are necessary for building up a number system? We recall that in English ten are used and all up to one hundred are but combinations of these ten. Of course two are sufficient — in a binary system. Many of the Eskimo tribes manage quite well with five. The Luiseno of California has but five abstract number words, but it has higher units which are chiefly descriptive phrases indicating various combinations of hands and feet. Many languages which have a decimal system get along easily without the full quota of ten as used in English. This is accomplished by the use of such combinations as $8 = 2 \times 4$, $6 = 2 \times 3$, 7, 8, 9 as 3, 2, 1 respectively subtracted from 10, and others which have already been given. The Coahuiltecan (given under the ternary system) forms all the numbers up to twenty with only four unique words, those for 1, 2, 4, 5 — a very remarkable instance.

3 Miscellaneous points

3.1 Limits in use

The numerals in some languages are given as high as a million and in many others to a thousand or more. We are naturally led to inquire whether such high numbers were actually in use by primitive people. The evidence found on this point cannot be here given in detail. Only a few conclusions may

be stated. The Eskimo seem to be poorest in ability to count, being low in comparison with the tribes of the United States. Most of them in ordinary conversation do not use above five or six, referring to higher numbers as “many,” but the more intelligent of them can count to 400. The Indians of the eastern United States who stand comparatively high intellectually, could use their numerals to 10,000 and their systems were such as to admit of indefinite expansion, one billion for instance being expressed as $1,000 \times 1,000 \times 1,000$. The Crees could count correctly as far as 1,000. The Winnebagoes are said to use their numerals as high as one million. Indefinite and countless numbers they represent by the terms “leaves on the trees,” “stars of the heavens,” “blades of grass on the prairie,” “sand on the lake shore.” The Crows do not count above a thousand, as they say honest people have no use for higher numerals! The Apaches cannot use numbers beyond 100,000. Any of the California tribes of which positive statements can be made can count into the hundreds. As a definite upper limit to counting ability we find that the Tuolomne are credited with the ability “to count with great rapidity almost to infinity”! [12, p. 406]

Speaking in general terms, we may say that the rather highly developed Siouan tribes of the Plains, the Iroquoian and Algonquian families of the east, and the Muskogean tribes of the South — all rather high in the scale of civilization — could count intelligently at least into hundreds of thousands and had words for even higher numbers; that most of the other Indians of the country had little actual knowledge of forms for numbers higher than thousands; while the Eskimo of the far north were limited to hundreds and in many cases to twenty or even ten. It is probable that before coming in contact with European civilization the Indians had little occasion to use numbers beyond a thousand. But the systems of many of them were such as to admit of indefinite and easy extension when needed.

3.2 Numeral classifiers

Some tribes use different sets of numerals for counting different classes of objects. The Tsimsian language has quite distinct forms for counting *men* and for counting *things*. More frequently the number stem is modified by a prefix or suffix which is in the nature of a classifier to denote the class of objects counted. The Haida has no less than 15 of these classifiers. Others seem to have an even larger number. This usage is found very generally among the

languages of the north Pacific coast and but rarely in other parts of the country. The Tsimsian mentioned above has different forms for abstract counting, for counting flat objects or animals, for round objects or time, for men, for long objects, for canoes, for measures.

3.3 Verbal nature

In a few languages the numerals are true verbs instead of adjectives, and as such are conjugated through all the variations of mood, tense, person and number. As far as found this peculiarity is limited to three languages, Cree, Crow, Micmac.

3.4 Derivative numerals

The formation of ordinals, adverbials and distributives from the cardinal numerals is more properly a grammatical than a numerical process and as such need receive only slight notice here.

(a) *Ordinals* are found most frequently, usually being formed from the cardinals by a suffix or other terminal modification, occasionally by a prefix. In the Creek an unusual method is found, the ordinal being formed from the cardinal in the same way that the superlative of the adjective is formed from the comparative.

(b) *Distributives* are also frequently found. They are formed from the cardinals by prefixes, suffixes, or reduplication.

(c) *Adverbials*, as far as noted, are formed by suffixes.

3.5 Arithmetical operations

We shall close our discussion of the varied, interesting, and intricate number systems of the North American Indians with a reference to their ability to perform arithmetical operations. In our study of the principles of formation of number words we found an extensive use of addition and multiplication, a less use of subtraction, and very slight use of division in the formation of number systems. But aside from these instances (all operations on only the *bases* of the systems) the calculative ability of the American Indians was very slight and of the most elementary sort. Addition, subtraction or multiplication was accomplished only with the aid of the fingers, sticks, pebbles or other convenient counters. It is probably that the native Indian mind had practically no idea of mental arithmetic, being unable to multiply or divide

numbers mentally, or even to add or subtract any except the smallest. His need for such operations was probably as slight as his knowledge.

4 Notes

4.1 Bibliographical note

It is impossible to give a comprehensive bibliography of this subject in a small space. Most of the material is in hundreds of separate vocabularies, grammars, dictionaries and discussions of the various languages of the American Indians. Considerable information may be found in the reports of the Bureau of American Ethnology, Washington, D.C. Pilling's *Bibliographies*, published by this Bureau from 1887 to 1894, contain references to most of the literature up to the date of their publication for the nine most important linguistic families. In addition, [2], [4], [17], and [18] may be mentioned as especially important. Full credit can scarcely be given for each statement made in this paper. The above mentioned sources have been used freely, but even more the numerous dictionaries and grammars mentioned at the beginning of this note. A bibliography of about 300 titles prepared by the author is on file in the library of the University of Chicago, the Newberry Library, Chicago, and the library of the University of Wisconsin.

4.2 Nature of sources

Before the arrival of the white man the Indian had no written language, except a system of rude hieroglyphics among some of the more intelligent tribes. Reproductions of many of these are given by Mallery in the 4th and 10th Reports of the Bureau of Ethnology. They have but little of mathematical interest. The oral number systems of the Indians are the important sources available. But these have been committed to writing by hundreds of different men, of varying reliability and familiarity with the languages, of different nationalities, using various systems of spelling, at different dates, and extending to various limits. Aside from this is the fact that the same Indians are often known by a dozen or twenty different names, or the same name is applied to several distinct tribes speaking different languages. Such confusion in the sources makes it extremely difficult to classify them satisfactorily for the purposes of comparative study. Many errors in detail must have been made which can only be corrected by further study

and reference to experts in American linguistics in various parts of the country. But it is hoped that such errors as have been made are not serious enough to affect much, if at all, the general results and statements made in this paper.

4.3 Nature of conclusions

The conclusions stated in this paper are usually based entirely on the relative *number* of the 324 examined. Little effort has been made to indicate the amount of territory covered or the number of Indians involved. Some of the languages were used by only a few people, others by many thousands. But for the purposes of this paper a small tribe with peculiarities in its number system is as interesting and as important as a much larger one.

References

1. Moritz Cantor, *Vorlesungen über Geschichte der Mathematik*, Leipzig: Teubner, 1880-1908.
2. L. L. Conant, *The Number Concept*, New York, 1896.
3. J. B. Dunbar, *The Pawnee Language*, 1893. (Pamphlet, copy in Newberry Library, Chicago)
4. R. B. Dixon and A. L. Kroeber, Numeral systems of languages of California, *American Anthropologist* 9.
5. D. W. Eakins, in Schoolcraft's *Indian Tribes*, Philadelphia, 1851-60.
6. K. Fink, *History of Mathematics*, Beman and Smith, trans., Chicago, 1900.
7. A. Gallatin, *Transactions of the American Ethnological Society*, 1 (1845).
8. A. S. Gatschet, Indian Numeral Adjectives, *American Antiquarian*, 2.
9. Grimm, *Geschichte der deutschen Sprache*.
10. H. Hankel, *Zur Geschichte der Mathematik in Alterthum und Mittelalter*, Leipzig, 1874.
11. C. Harrison, Haida Grammar, *Proceedings and Transactions of Royal Society of Canada*, (2nd series), 1, sec. 2.
12. A. Johnson, in Schoolcraft's *Indian Tribes*, Philadelphia, 1851-60.
13. J. Johnson, Present State of Indian Tribes Inhabiting Ohio, *Proceedings of the American Antiquarian Society* 1.
14. E. W. Nelson, Eskimo about Bering Strait, in *Eighth Annual Report, Bureau of Ethnology*, 1896-97.
15. P. Petitot, *Vocabulaire Francaise Esquimau*, Paris, 1876.
16. J. W. Powell, *First Annual Report, Bureau of Ethnology*, 1879-80.
17. ———, Linguistic Families of America, North of Mexico, in *Seventh Annual Report, Bureau of Ethnology*, 1885-86.
18. J. H. Trumbull, On Numerals in American Indian Languages, *Transactions of the American Philological Association*, 1874.
19. D. Zeisberger, Grammar of the Delaware, *Transactions of the American Philosophical Society*, 3.

The Number System of the Mayas

A. W. RICHESON

American Mathematical Monthly 40 (1933), 542–546

The number systems of the North American Indians have recently been discussed in detail in two papers in this *Monthly* [2]. The system of numbers developed by the semi-civilized Maya Indians of Central America is probably the most interesting of all systems developed by the early inhabitants of this continent.

The examples of the number system of the Mayas that have been found, or at least that have been deciphered, deal with the counting of time events or periods, and many authorities are of the opinion that the recording of time series was the sole purpose of their numbers. The records of their chronicles are found as glyphs on the monuments and as written in the codices. These records present two methods of writing numerals, the normal form and the head-variant form. Both forms are essentially the same, and the Mayas were able to express a number as easily by one method as by the other. The head-variant form is found with few exceptions on the monuments, while the normal form is found exclusively in the codices.

In the head-variant form there are distinctive head forms for each of the numbers from 0 to 12 inclusive, while from 13 to 19 inclusive the numbers are written by using the head form for 10 plus the form for whatever unit is needed to make up the desired number. Each number is characterized by a distinctive type of head, by means of which it can be distinguished from any other number. In the case of three numbers, 2, 11, and 12, however, the characteristic elements have not been determined with certainty. The forms for these numerals occur very rarely on the inscriptions, and consequently, the data are not sufficient to justify a statement as to the characteristic elements.

Figures 1-3 illustrate the head forms for 6, 10, and 16 respectively. Figure 4 shows the head form for 16



Figure 1.



Figure 2.



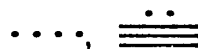
Figure 3.



Figure 4.

used as a multiplier with the kin or day sign to the right. It should be noted that the head form for 16 is made up of the fleshless jaws of the character for 10 with the “hatchet” eye for 6. The characteristic elements for the numbers from 0 to 19 are given in the table on the next page.

In the normal form the number combinations from 1 to 19 inclusive are formed by dots and bars. Each dot has the numerical value of 1 and each bar represents five. Generally the dots are placed horizontally over the bars or to the side of a vertical arrangement of the bars; for example, 4 and 17 were written respectively as follows.



On the inscriptions the number forms were frequently decorated to give them symmetry and a balanced form; this has often been a source of error in deciphering the inscriptions. Since the Mayas used a vigesimal system of numeration, there was no need of a symbol for twenty, since 20 units of the first

<i>Head form</i>	<i>Characteristic element</i>
0	Clasped hands across lower part of face
1	Forehead ornament composed of more than one part
2	Undetermined
3	Banded head dress
4	Bulging eye with square irid, snag tooth, curling fangs from back of mouth
5	“Tun” sign for head dress
6	Hatchet eye
7	Large scroll passing under eye and curling under forehead
8	Forehead ornament composed of one part
9	Dots on lower cheek or around mouth
10	Fleshless lower and upper jaws
11	Undetermined
12	Undetermined—type of head known
13 to 19	Head for 3, 4, 5, 6, 7, 8, 9 with fleshless lower jaw for 10

20 tuns = 1 katun = 7,200 days
 20 katuns = 1 baktun = 144,000 days
 etc.

On the other hand, Dr. Teeple of the Carnegie Institute and Mr Wm. E. Gates of The Johns Hopkins University have advanced the opinion that the tun is the correct unit of time used by the Mayas, and that the time count is vigesimal throughout. They argue that the division of the tun or year into 18 and 20 parts is nothing more than fractional parts of the Maya time unit [10].

Figure 5 illustrates the formation of the number 75,550 on the basis of the above table by employing the dot and bar characters for the multipliers with the kin, uinal, tun, and katun signs as multiplicands. Reading from the top down, we have 10 katuns = 72,000 kins, 9 tuns = 3,240 kins, 15 uinals = 300 kins and 10 kins = 10 kins. The sum of the four products is 75,550 kins. This number could be expressed also by the head forms for the multipliers 10, 9, 15, and 10, in place of the dot and bar characters.

The second method of numeration, namely that by position, is very similar to the Hindu-Arabic decimal system as used today. Although the system is vigesimal and thus required 19 different combinations for the units, it was built up by the three simple characters: the dot, the bar, and the zero. With this method the Mayas were forced to fix arbitrarily a starting point and to confine themselves to one series only, or else the positional value of the nineteen digits would be useless. They accordingly adopted an ascending

order gave one unit of the second. However, a symbol for zero was absolutely indispensable, and this symbol, which somewhat resembled the shape of a shell, is found on the inscriptions and in the codices. The symbol was first recognized by Dr. Förstemann [4].

1 Methods of numeration

The Mayas developed two systems of numeration; the multiplication method and the “numeration by position”. Although different in form, both methods are essentially vigesimal.

The first method, which is rarely found except on the inscriptions, makes use of both the normal and head-variant forms. The numbers are formed by using the bar and dot characters or the desired head form to build up the multipliers from 0 to 19 inclusive, with the time period signs as multiplicands. Until recently most authorities have stated that the Maya time count was one of days or kins and that the count was not strictly vigesimal. The following table will show the count under this assumption:

1 kin = 1 day
 20 kins = 1 uinal = 20 days
 18 uinals = 1 tun = 360 days

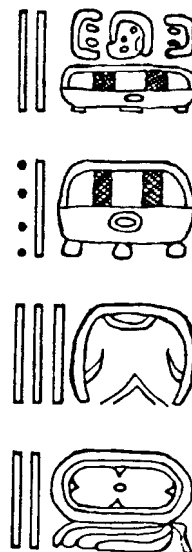


Figure 5.

$$\begin{array}{r}
 \cdot \cdot \cdot \cdot = 4 \times 2,880,000 = 11,520,000 \\
 \overline{\cdot} = 6 \times 144,000 = 864,000 \\
 \overline{\cdot \cdot \cdot \cdot} = 14 \times 7,200 = 100,800 \\
 \overline{\cdot \cdot \cdot} = 13 \times 360 = 4,680 \\
 \overline{\overline{\cdot \cdot \cdot}} = 15 \times 20 = 300 \\
 \cdot = 1 \times 1 = 1 \\
 \hline
 12,489,781
 \end{array}$$

Figure 6.

series which corresponds to our decimal series from right to left.

We illustrate in Figure 6 the method by the number 12,489,781. This is the largest number yet found in the codices.

2 Discussion of the numbers

The Maya numbers were no doubt written as they were spoken. The names of the numbers from 1 to 20 inclusive are given below as they appear in Beltran’s *Arte del Idioma Maya*.

1 hun	11 buluc
2 ca	12 lahca
3 ox	13 oxlahun
4 can	14 canlahun
5 ho	15 lolahun
6 uac	16 uaclahun
7 uuc	17 uuclahun
8 uaxac	18 uaxaclahun
9 bolon	19 bolonlahun
10 lahun	20 hunkal or kal

Very little seems to be known concerning the origin of the number words from 1 to 5 inclusive. Dr. Brinton [1] is of the opinion that the Maya proper and the neighboring Mayan dialects were derived from one common archaic form of speech and not from one another. In the case of the smaller numbers, this opinion seems to be justified, as they were no doubt formed before the beginning of their history. A number of arguments have been advanced for the derivation of the numbers from 5 to 9. Dr. Thomas [11] believes that the hand was not used in the count until 5 was reached and that the numbers from 6 to 8 inclusive were composite. He suggests that *uaxac* is the answer to the whole question, that the *x* of *ox* = 3 has been combined by *u* with some form of 5 to give eight and that the forms for 6 and 7 are formed in a similar manner. Pio Perez on the other hand gives as a signification of the verb *uac*

or *uach* “to take out one thing which is placed in another and united with it.” This would seem to indicate counting on the fingers and turning them in for the first five and then opening them out while counting the second five. *Bolon* = 9 seems to have the meaning “on the way to 10”, while *lahun* = 10 is *lah hun*; “it finishes one man,” i.e., counting on the fingers.

The numbers from 12 to 19 inclusive are without doubt composite numbers, i.e., 12 = 10 + 2, 13 = 10 + 3, etc. As we should expect in a vigesimal system, there is a definite number for 20, *kal* or *hunkal*. Henderson gives for *kal* “to close, to shut” or as a substantive “a fastening together”, i.e., a fastening together of both hands and feet.

The count from 21 to 40 inclusive is by addition to the first 20, e.g., 21 is *hun-tu-kal* = 1 + 20 or 1 to the 20. Forty is *ca-kal* or 2 × 20. From 41 on, the count is regular, but is different from 21 to 40. Here the count is by subtraction from the next 20 [8].

Dr. Brinton states that the Maya’s use of numbers was somewhat different from ours [1]. The numbers are rarely used except with a numeral particle, which is suffixed to the numeral and indicates the character or class of the objects which are about to be enumerated. With the aid of these particles Dr. Brinton gives another method which was frequently employed to express their numbers. For eighty-one years they did not write *hun tu yox kal haab*, as we would expect, but *can kal haab catac hunpl haab*, i.e., four score years and one year.

3 Conclusion

It is quite evident that the Mayas had developed a number system with a place value for their characters many years before the advent of the white man. Dr. Teeple is of the opinion that the Mayan vigesimal system of numbers was a distinct part of an American civilization [10].

The records also indicate that the Mayas were unable to handle fractions as we do today, but on the other hand they were able to and did perform long numerical computations involving multiplication and division. Just how these computations were carried out we do not know. Dr. Förstemann in his work gives an instance from the inscriptions where the calculation runs into the millions [3].

It is impossible to give complete references for many of the statements, but the material has been drawn largely from the following sources.

References

1. Dr. Brinton, *Maya Chronicles*.
2. W. C. Eells, Number Systems of the North American Indians, *American Mathematical Monthly*, 20 (1913), 263–279; 292–299.
3. Förstemann, *Zur Entzifferung der Maya-Handschriften*, No. II.
4. —, *Zur Maya Chronologie*, *Zeitschrift für Ethnologie*, (Berlin) 1891.
5. W. J. McGee, Primitive Numbers, *19th Annual Report of the Bureau of American Ethnology*, 821–851.
6. G. Morley, *An Introduction to the Study of Maya Hieroglyphs*, Bureau of American Ethnology, Bulletin 57.
7. S. G. Morley, *The Inscriptions of Copan*, Carnegie Institution, Washington, D. C., 1920.
8. Rosney, Mémoire sur la numération dans la langue et dans l'écriture sacrée des anciens Mayas, *Compte-Rendu de Congrès International des Américanistes*, (Paris 1875), vol. 2, p. 439.
9. Dr. Spinden, *Ancient Civilizations of Mexico*.
10. John Teeple, *Maya Astronomy*, No. 11, Carnegie Institution, Washington, D. C.
11. Cyrus Thomas, Numerical Systems of Central America and Mexico, *19th Annual Report of the Bureau of American Ethnology 1897-98*, 853–955.

Before The Conquest

MARCIA ASCHER

Mathematics Magazine 65 (1992), 211–218

1 Introduction

In the late 15th century, through their explorers, Europeans “discovered” the New World. Although the discovery would cause drastic change, the New World was, of course, not new to its inhabitants. When the Europeans arrived, there were at least 9 million people in about 800 different cultures living in the Western Hemisphere. Because of the vast disruptions that eventually took place, what we know about them and their mathematical ideas is limited. Most of the cultures had no writing as we commonly use the term and so there are no writings by them in their own words. For the cultures that did not survive, we have primarily what can be learned from archaeology and from the writings of the Europeans of the time, who had little understanding and little respect for these cultures so different from their own. For those that did survive, we also have their oral traditions.

We focus here on the mathematical ideas of two sizable groups, the Incas and the Maya. The regions the groups inhabited, their cultures, and their histories are quite distinct, as are their mathematical ideas. Fortunately, for both groups, there is sufficient information for us to gain some understanding of their rich and complex ideas. Here we present an abbreviated introduction to the content and context of the sophisticated data handling system of the Incas and the intricate calendric system of the Maya.

2 The Incas

The Incas comprised a complex state of about 5 million people that existed from about 1400 C.E. to 1560 C.E. in what is now Peru, and also parts of modern Ecuador, Bolivia, Chile, and Argentina. There were

many different peoples in the region, but, starting about 1400, the Incas forcibly consolidated the others into a single bureaucratic entity. The consolidation was achieved by the overlay of a common state religion and a common language, relocation of groups of people, extensive systems of roads and irrigation, and a system of taxation involving, for example, agricultural products, labor, and cloth. The Incas also built a network of storehouses to hold and redistribute goods as well as to feed the army as it moved. The Inca bureaucracy can be characterized as methodical, highly organized, and intensive data users. Although the Incas did not have what we call writing, they did keep extensive records. These were encoded via a logical-numerical system on spatial arrays of colored, knotted cords called *quipus*.

A few selected people from each region that the Incas occupied were trained to serve in the Inca administration and, in particular, to be responsible for gathering, and then encoding and decoding a wide variety of information on the quipu. Believing the quipu to be works of the Devil, the Spanish destroyed thousands of them. Only about 500 remain. These were recovered from graves, probably buried with those who made them. Only rarely can we read the quipu in the sense that specific meaning can be assigned. However, we can reconstruct something of their logical-numerical system and, as a result, see the interrelationships of some of the data they contain.

A photograph of a quipu is in Figure 1. Figure 2 is a schematic. In general, a quipu has a main cord from which other cords are suspended. Most of the suspended cords are attached such that they fall in one direction (*pendant cords*); some few fall in the opposite direction (*top cords*). *Subsidiary cords* are often suspended from the pendant or top cords. And

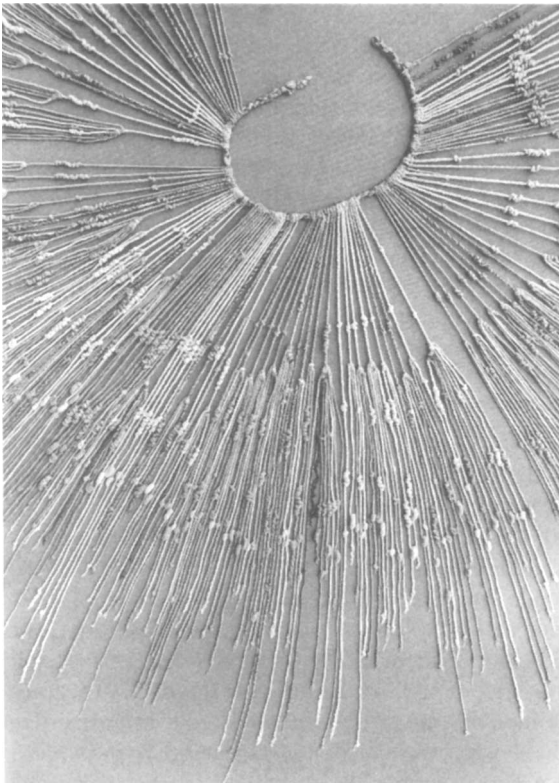


Figure 1. A quipu in the collection of the Museo Nacional y Arqueología, Lima, Peru. (Photo by Marcia and Robert Ascher.)

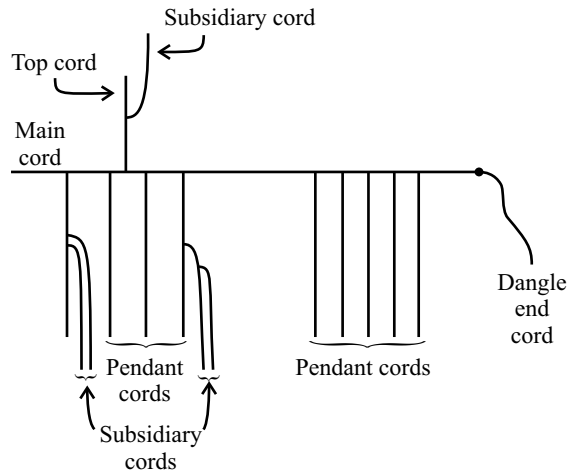


Figure 2. A schematic of a quipu.

there can be subsidiaries of subsidiaries, and so on. (Notice that in Figure 2 the first pendant has two subsidiaries on the same level while the fourth pendant has two levels of subsidiaries.) Some pendant cords have as many as 18 subsidiaries on one level, and some have as many as 10 levels of subsidiaries. Sometimes a single cord (*dangle end cord*) is attached to the end of the main cord in a way that sets it apart from the pendant and top cords. All cord attachments are tight so that the spacing between the cords is fixed and serves to group or separate the cords. Overall, a quipu can be made up of as few as three cords or as many as 2000.

Color is also a feature of the logical system. It is used primarily to associate or differentiate cords within a single quipu. Thus, color as well as space can create cord groupings. For example, eight pendants can be formed into two groups by having four white pendants followed by four green pendants, or by a four-color sequence repeated twice. In the latter case, each cord is not only associated with its group, but also with the like-colored cord in the other

group. Similarly, subsidiaries are associated or differentiated by color as well as by level and relative position on the given level.

Spaced clusters of knots on the cords represent numbers. No matter what the cord placement, only three types of knots appear (single knots, long knots, and figure-8 knots). Depending on the knot types and relative cluster positions, each cord can be interpreted as one number or as multiple numbers. If it is one number, it is an integer in the base 10 positional system. Each knot cluster is read as a digit and each consecutive cluster, starting from the free end of the cord, is valued at one higher power of 10. The units position is always a long-knot cluster or a figure-8 knot, while all other positions are clusters of single knots. When, instead, the cord carries multiple numbers, long-knot clusters or figure-8 knots are interspersed with single-knot clusters thereby signaling the start of a new number. The color coding of the cords also helps in the interpretation of values by enabling the distinction between a numerical value of zero and an intentional omission or blank.

Knot types and knot positions, cord directions, cord levels, color, and spacing are all structural indicators that were combined together in sufficiently standardized ways to be read and interpreted by the community of quipumakers. That is, the quipus served for communication, not as ad hoc personal mnemonic devices. Top cords, for example, generally carry the sum of the pendant cords with which they are grouped on the main cord. Another aspect of the system that is crucial to its general applicability is that numbers were used as labels as well as magnitudes. Particularly with the advent of computers,

this usage is now very prevalent in our own culture. For example, the composite number 202-387-5200 is a label identifying a geographic region, a locale within that region, and a specific telephone within that locale.

The quipus, then, are logically structured arrays of magnitudes and labels. Let us translate three of them into notation that is familiar but preserves their logical structure. Then we can delve into some of their internal data relationships.

The quipu shown in Figure 1 contains solely quantitative data. Analysis of the pendant cord colors and spacing shows that there are six ordered sets of 18 values each. We will call the j th value in the i th set a_{ij} , where $i = 1, \dots, 6; j = 1, \dots, 18$. When the knots are interpreted as magnitudes, we find that, for all j ,

$$a_{1j} = a_{2j} + a_{3j}$$

and, in turn,

$$a_{2j} = \sum_{i=4}^6 a_{ij}$$

Hence, the relationship

$$a_{1j} = \sum_{i=3}^6 a_{ij}$$

also holds. Additionally, there are subsidiary cords on the pendants in five of the six cord groups. Thus, for each a_{ij} , for $i = 2, \dots, 6; j = 1, \dots, 18$, there are as many as 11 ordered subsidiary values. Call them

$$a_{ijk}, \text{ with } k = 1, \dots, 11.$$

Here, too, consistent summation relationships exist:

$$a_{2jk} = \sum_{i=3}^6 a_{ijk} \text{ for } k = 1, \dots, 11; j = 1, \dots, 18.$$

A modern analogy of data with sums of sums and sets of sums, as is seen in this example, is an accounting scheme for a company, broken down to reflect that it is made up of several departments and producing a variety of products.

In our second example, the arrangement of values and their sums is analogous to a matrix that has, as a subset, the transpose of the sum of two other matrices. Specifically, this quipu's data can be thought of as two 3×3 matrices, each preceded by a single value, and a 3×5 matrix. Calling the elements of the matrices

$$a_{ijk} \quad i = 1, 2, 3; j = 1, 2, 3; k = 1, 2,$$

and

$$b_{ij} \quad i = 1, 2, 3; j = 1, 2, 3, 4, 5,$$

the relationship is

$$b_{i,2j-1} = \sum_{k=1}^2 a_{ijk} \quad \text{for } i = 1, 2, 3; j = 1, 2, 3.$$

And, continuing the analogy to matrices, the single value preceding each of the 3×3 's would be the sum of its first row; that is,

$$c_k = \sum_{j=1}^2 a_{1jk} \quad \text{for } k = 1, 2.$$

Some of the data structures remind us of spreadsheets, matrices, and tree diagrams. Other quipus have other layouts, non-quantitative as well as quantitative data, other kinds of internal data relationships, or even relationships with data on other quipus. Many remain fascinating puzzles. One of these, our final example, is from a pair of quipus that were found together.

The specific numbers on these two quipus are different, but the quipus share several internal data relationships, including what we commonly call a difference table. While both appear to be expressions of the same algorithm, a concise unifying description escapes me. Translated into tabular form, they are compared in Figure 3. I have superimposed arrows and heavy lines on the tables to indicate the similarities that I see. Perhaps you can find additional similarities or, perhaps, you can find a generalization that unites the data sets.

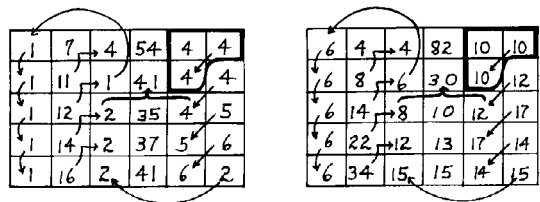


Figure 3. Data excerpted from a pair of quipus found together. Arrows and heavy lines highlight some of their similarities. In both, for example, all values are the same in the first column and in row 2, column 3. Also, in both, the third column contains the differences of consecutive values in the second column. (The quipus are described by C. Radicati di Primeglio in “La ‘seriación’ como posible clave para descifrar los quipus extranumerales”, *Documenta: Revista de La Sociedad Pemana de Historia* 4 (1965), 112–215.)

Reference [1] contains more details, examples, and discussion of the context, contents, and interpretation of quipus. Although we lack the cultural associations needed to know what a specific quipu means, the quipus surely are records of human and material resources and calendric information. But they contain much more — possibly information as diverse as construction plans, dance patterns, and even aspects of Inca history. Overall, the logical-numerical system embedded in these spatial arrays of colored, knotted cords was sufficiently *general* to serve the needs of the Inca bureaucracy. Their use was terminated soon after the destruction of the Inca state in 1560 C.E.

3 The Maya

The Mayan peoples have a complex cultural tradition extending over a long period of time and encompassing different groups speaking about 25 different languages. They shared much in the way of culture but, spread through time and space, they had different centers and political organizations, some different ideas, and some different practices. Discussions of their history usually begin sometime before 1000 B.C.E. The period 200–1000 C.E. is referred to as the Classic period and is marked by ceremonial centers with monumental architecture, a system of writing, an elaborate astrological science, and numerous centers of social, religious, economic, and political activities interrelated by marriage and trade networks. During the Classic period, the Maya inhabited what are now the eastern Mexican states of Chiapas, Tabasco, Campeche, Quintano Roo, and Yucatan; Belize; Guatemala; and the western portions of Honduras and El Salvador. On the eve of the Spanish conquest, there were spread out in this area, many independent yet culturally interrelated states, none as grandiose as earlier. Because they were dispersed and independent, they did not succumb to the Spanish as quickly and easily as did the Incas. Today, primarily in Chiapas and the highlands of Guatemala, some Maya traditions continue.

Christopher Columbus, in 1502, is said to have been the first European to encounter the Maya, and his brother, Bartholomew, was the first to record the name of the group. By that time, however, remains of the Classic Maya period were already covered over, and so another “discovery” — this time archeological — took place beginning in the mid-1800s. In addition to some ongoing traditions, what we know

of the Maya, and in particular of their mathematical ideas, comes from archeological materials, including thousands of inscribed stone monuments (*stelae*) and four post-Classic books (*codices*), the only ones remaining of the *thousands* that were burned by the Spanish.

We will concentrate on the idea of *time* as it permeates the Mayan culture. Time is considered to be cyclic. Supernatural forces and beings are associated with and influence units of time. Events of the past, present, and future are related through the recurrence of named time units. There are, however, not just one, but several, overlapping cycles that all must be taken into consideration to give meaning to any particular time unit. Although their calendric concerns extend to the incorporation of astronomical phenomena, the Maya were preoccupied with the interrelationship of the arbitrary cycles they created and imposed on time. For this reason, the Maya are said to have “mathematized” time and, through it, their religion and cosmology.

There is, first of all, a 260-day *ritual almanac*. Each day within it is identified by a number in a cycle of 13 and a named deity in a cycle of 20. (Each of the 13 numbers also has an associated deity.) There is a *vague year* of 365 days (called “vague” because it does not keep in alignment with the seasons). It results from a cycle of 20 numbers *within* a cycle of 18 named deities plus five unnamed days. (The cycle of 20 is now referred to as a month but does not have a lunar correspondence.) One *calendar round* is 18,980 days (52 vague years, 78 almanac cycles) since that is the least common multiple of 365 and 260. A date within this, made up of an almanac date and a vague year date, reads, for example, 4 Ahau 8 Cumku where Ahau and Cumku are names of deities.

In the ceremonial centers of the Classic period, there were temples atop large, stepped pyramid frusta as high as 213 feet. Hundreds of *stelae*, some as tall as 32 feet, were erected around them to commemorate different events. To mark an event, what was needed was to accurately and *fully* identify it in time and, sometimes, to state how many days it was from another event. In addition to the calendar round date, another significant identifier was a *Long Count*: the number of days from the beginning of the then current *Great Cycle*. A Great Cycle is based on a 360-day period (a *tun*) consisting of 18 *uinals* of 20 days each; 20 *tuns* are a *katun*, 20 *katuns* are a *baktun*, and 13 *baktuns* are a Great Cycle. An example of a Long Count transcribed into our numerals is

9.0.19.2.4. From left to right this reads “9 baktuns, 0 katuns, 19 tuns, 2 uinals, and 4 days.” To convert to our system, starting at the right, each position—with the exception of the third—is multiplied by one higher power of 20. In the third position, an 18 is used instead. Hence, the Long Count date of 9.0.19.2.4 is interpreted as:

$$9 \cdot 18 \cdot 20^3 + 0 \cdot 18 \cdot 20^2 + 19 \cdot 18 \cdot 20 + 2 \cdot 20 + 4 \\ = 1,302,884 \text{ days}$$

from the beginning of the Great Cycle that started on the calendar round date of 4 Ahau 8 Cumku. The exact correlation of this date with the Gregorian calendar is not known. However, by one commonly accepted correlation, the beginning of the Great Cycle was in 3114 B.C.E. and the date given by the Long Count above is, thus, in 454 C.E.

This Long Count date appeared on a stela that also was dated in the calendar round as 2 Kan 2 Yax. But, as with most stelae, it had dates placing it within still more cycles. There was a 9-day cycle of *Lords of the Night*, each associated with one of the nine levels of the underworld. Hence, a specific Lord of the Night also dates the day being marked. And, in addition, the day is placed within a *lunar* cycle. Lunar years and half-years are made up of 29- and 30-day lunar months. The stela contains the moon number within the lunar half-year, the age of the moon, and whether it is a 29- or 30-day month.

Just as there are nine levels below the earth, there are 13 levels in the heavens above. There are four cardinal directions and each of the quadrants they define is associated with a different color. Uniting time and space, the days of the 260-day ritual almanac move in a counterclockwise direction through the four quadrants. Hence, not only are time and space related, but the ritual almanac has within it a four-color cycle. In some cases, where dates also identify days within a 819-day cycle associated with the rain god, the use of four colors effectively makes that cycle $819 \cdot 4 = 3276$ days.

Many of the Maya computations are projections into the past or into the future that require dovetailing the cycles. For instance, one inscription, commemorating the enthronement of a ruler, gives the calendar round dates of his birth and his enthronement, as well as of the enthronement of an earlier, somehow related ruler or deity. The number of days between these events is also included in Long Count form. For example, the time elapsed since the enthronement of the deity is 7.18.2.9.2.12.1 days.

Hence, given one calendar round date, a calendar round date some $1\frac{1}{4}$ million years earlier was calculated or, given two calendar round dates, their Long Count difference (number of days between them) was calculated.

To more fully savor the calculation, you might try to do such a problem. First, recall that each of the 260 days in the ritual almanac is identified by a number in a cycle of 1 to 13 and a named deity in a cycle of 20. For simplicity, let us call the deities D_1, \dots, D_{20} . In the 365-day vague year, five days are unnamed while, for the rest, each day is identified by a number in a cycle of 1 to 20 within each of 18 months named for deities. Call these deities d_1, \dots, d_{18} and assume that the five unnamed days follow $20d_{18}$. The calendar round date is the almanac date followed by the vague year date. What, then, is the calendar round date that is 2.3.5.10 days after $8D_{10}13d_{10}$? And, what is the Long Count Difference between $12D_86d_2$ and the next $5D_4l2d_{17}$? (Answers at the end of the article.)

The Dresden Codex, attributed to the eleventh century in Yucatan is the most mathematical of the codices. It is constructed as a long strip of paper made from tree bark, folded into pages, coated with white plaster, and painted. Perhaps as aids to computation, the codex includes several tables of multiples; for example, there are tables of multiples of 5.1.0 (that is, of 1820 that equals $7 \cdot 260$ and $5 \cdot 364$) up to 1.0.4.8.0. But, even more important, other tables in the codex combine backward and forward calendar projections with evidence of keen astronomical knowledge. One set of tables correlates lunar cycles with ritual almanac dates. These tables cover 405 lunations and are interpreted as prediction tables for possible eclipses. Another set of tables in this codex correlates Venus visibility events with ritual almanac and vague year dates. Covering 65 Venus cycles, which is 146 almanac cycles and 104 vague years, it includes corrections reflecting the fact that the mean synodic year of Venus is not an integral number of days. (The mean synodic year of Venus is 583.92 days.) The corrections are such that the error between real and tabulated times of the positions of Venus would be off by just two hours in about 500 years!

We know that much is unknown about the knowledge and mathematical ideas of the Maya. Dates and numbers, written in a variety of symbolic forms, have been recognized and deciphered. But the Maya writings contain much more. The writing system is complex because it contains about 1000 symbols and

both phonetic and non-phonetic elements. A great deal of recent activity in decipherment raises the hope that more will become known about Maya ideas and the Maya culture in general. (For more details on number representation, the tables in the Dresden Codex, and possible algorithms for the date difference calculations, see [2] and [5]. Reference [3] discusses the importance of the Maya scribes including evidence that they were both women and men. Also, [4] is an excellent comprehensive overview of the Maya.)

4 Conclusion

The Inca and Maya are two substantial examples of cultures whose mathematical ideas were both sophisticated and independent of those of Western culture. We can never know about all of the mathematical ideas they had and, what is more, we cannot know what they might have developed had they continued to thrive.

Answers to Exercises: $11D_{20}8d_5$; 1.19.6.16.

References

1. M. Ascher and R. Ascher, *Code of the Quipu. A Study in Media, Mathematics and Culture*, University of Michigan Press, Ann Arbor, MI, 1981.
2. M. Closs, The mathematical notation of the ancient Maya, *Native American Mathematics*, ed. by M. Closs, University of Texas Press, Austin, TX, 1986, pp. 291–369.
3. —, *My mother and my father were both scribes*, Research Reports on Ancient Maya Writing, Center for Maya Research, Washington, DC, in press.
4. J. S. Henderson, *The World of the Ancient Maya*, Cornell University Press, Ithaca, NY, 1981.
5. F. G. Lounsbury, Maya numeration, computation, and calendrical astronomy, *Dictionary of Scientific Biography*, Vol. 15, suppl. 1, Scribners, New York, 1978, pp. 759–818.

Afterword

The two standard accounts of Mesopotamian mathematics (as well as the mathematics of other ancient civilizations) are Otto Neugebauer's *The Exact Sciences in Antiquity* [14] and B. L. Van der Waerden's *Science Awakening I* [16]. Although they are both still useful, they have been superseded in some of their technical accounts of the mathematics by the results of new research. Among the newer surveys of Mesopotamian mathematics are articles by Jens Høyrup [7] and Jöran Friberg [5]. Høyrup also has a book-length treatment of the technical aspects of the Mesopotamian tablets: *Lengths, Widths, Surfaces: A Portrait of Old Babylonian Algebra and Its Kin* [9] as well as a series of more general essays on ancient and medieval mathematics: *In Measure, Number, and Weight: Studies in Mathematics and Culture* [8].

The standard, and still useful, history of Greek mathematics is Thomas Heath's *A History of Greek Mathematics* [6]. But many aspects of Heath's analysis have been challenged in recent years. The two best reevaluations of some central parts of the story of Greek mathematics are Wilbur Knorr's *The Ancient Tradition of Geometric Problems* [10], which argues that geometric problem solving was the motivating factor for much of Greek mathematics, and David Fowler's *The Mathematics of Plato's Academy: A New Reconstruction* [4], which claims that the idea of anthyphairesis (reciprocal subtraction) provides much of the impetus for the Greek development of the ideas of ratio and proportion. A newer work, Serafina Cuomo's *Ancient Mathematics* [3], is an excellent survey of Greek mathematics, aimed particularly at non-specialists.

There are now two good surveys of the history of Chinese mathematics available in English: *Chinese Mathematics: A Concise History* [11], by Li Yan and Du Shiran and *A History of Chinese Mathematics* [13] by Jean-Claude Martzloff. In addition, there is now a complete English translation, with commentary, of the classic *Nine Chapters on the Mathematical Art* [15].

For more information on Peruvian *quipus*, the best source is *Mathematics of the Incas: Code of the Quipu* [1] by Marcia and Robert Ascher. This book provides a mathematical analysis of various techniques of *quipu* making and also provides exercises for students. A good survey of Mayan mathematics is Floyd Lounsbury's article [2] in the *Dictionary of Scientific Biography*. Finally, *Native American Mathematics* [2], edited by Michael Closs, provides more up-to-date information on the number systems of North American Indians, as well as material on the mathematics of other Native American civilizations.

References

1. Marcia and Robert Ascher, *Mathematics of the Incas: Code of the Quipu*, Dover, New York, 1997.
2. Michael P. Closs, ed., *Native American Mathematics*, University of Texas Press, Austin, 1986.
3. Serafina Cuomo, *Ancient Mathematics*, Routledge, London, 2001.

4. David H. Fowler, *The Mathematics of Plato's Academy: A New Reconstruction*, Oxford University Press, 1999.
5. Jöran Friberg, *Mathematik, Reallexikon der Assyriologie* 7 (1987–1990), 531–585 (in English).
6. Thomas Heath, *A History of Greek Mathematics*, Dover, New York, 1981 (reprinted from the 1921 edition).
7. Jens Høyrup, Mathematics, algebra, and geometry, in *The Anchor Bible Dictionary*, David N. Freedman, ed., Doubleday, New York, 1992.
8. ———, *In Measure, Number, and Weight: Studies in Mathematics and Culture*, State University of New York Press, Albany, 1994.
9. ———, *Lengths, Widths, Surfaces: A Portrait of Old Babylonian Algebra and Its Kin*, Springer, New York, 2002.
10. Wilbur Knorr, *The Ancient Tradition of Geometric Problems*, Birkhäuser, Boston, 1986.
11. Li Yan and Du Shiran, *Chinese Mathematics: A Concise History*, trans. by John N. Crossley and Anthony W.-C. Lun, Clarendon Press, Oxford, 1987.
12. Floyd G. Lounsbury, Maya Numeration, Computation, and Calendrical Astronomy, *Dictionary of Scientific Biography*, vol. 15, Scribners, New York, 1978, 759–818.
13. Jean-Claude Martzloff, *A History of Chinese Mathematics*, trans. by Stephen S. Wilson, Springer, Berlin, 1997.
14. Otto Neugebauer, *The Exact Sciences in Antiquity*, Princeton University Press, 1951.
15. Shen Kangshen, John N. Crossley and Anthony W.-C. Lun, *The Nine Chapters on the Mathematical Art: Companion and Commentary*, Oxford University Press, 1999.
16. B. L. Van der Waerden, *Science Awakening I*, Oxford University Press, New York, 1961.

A decorative frame with a double-line border and rounded corners, containing the title text.

**Medieval
and Renaissance
Mathematics**

Foreword

Although the Middle Ages are often thought of as a period of little progress in mathematics, the statement is true only of Europe; much progress was made in other parts of the world. The first three papers in this section deal with the contributions of medieval south Indian mathematicians to the development of the power series representation of the sine, cosine, and arctangent series; these power series first occur in a work by Nilakantha in the early sixteenth century. A detailed derivation of the series appeared later in that century in a work of Jyesthadeva, who attributed the series to the fourteenth-century mathematician Madhava. This Indian work was first brought to the attention of western scholars by C. M. Whish in 1835, but his work had no effect. They were reintroduced to Europe in a series of articles by C. Rajagopal and his associates beginning in 1949.

The article by Ranjan Roy discusses the derivation of the arctangent formula and its application to finding a series approximation to π . Roy also discusses the analogous work by Gottfried Leibniz around 1675 and by James Gregory a few years earlier. Victor Katz's article concentrates on the derivation of the sine and cosine series. Since it was necessary for the derivation of all three series for the Indian mathematicians to have some knowledge of formulas for the sum of integral powers, Katz discusses one particular derivation of such formulas. This was the work of Ibn al-Haytham, known to the West as Alhazen, a mathematician who worked in Egypt around the year 1000. Finally, David Bressoud looks at the question of finite differences and how the Indian knowledge of these helped lead to the sine series. He also outlines Narayana's fourteenth-century derivation of the sum formula for integral powers. Both Katz and Bressoud consider the question of how close Islamic and Indian mathematicians came to inventing the calculus.

In another article dealing with an Islamic mathematician, Farhad Riahi considers al-Kashi's derivation of the formula for the sine of a triple angle and its use in determining the sine of 1 degree to an arbitrary level of accuracy. This work was one of many in Islam in which polynomial equations were solved numerically.

The next two articles in this section deal with mathematics in medieval Europe. The most important European mathematician of medieval times was Leonardo of Pisa (now known as Fibonacci). Although he is most famous for his major work on arithmetic and algebra, R. B. McClenon describes in some detail his *Book of Squares*, a treatise on some aspects of number theory which was unequalled until the time of Euler. Leonardo was also one of the first to introduce the Hindu-Arabic decimal place-value system to Europe. Barbara Reynolds discusses the controversies surrounding the introduction of this system and the conflict between the 'modern' users of paper-and-pencil algorithms and the 'traditional' users of the abacus for calculation.

We next move to the Renaissance. The basic story of the discovery of the cubic formula in sixteenth-century Italy is well known. Martin Nordgaard discusses some of the delightful details

behind this story, based on his examination of the series of challenges and responses between Niccolo Tartaglia and Ludovico Ferrari. As we see from this article, progress in mathematics is sometimes accompanied by intrigue and conflict. To provide additional insight into Renaissance mathematics, Abraham Arcavi and Maxim Bruckheimer take you through a section of Rafael Bombelli's *Algebra*, in which they extract a square root. The authors show the relationship of Bombelli's method to the method of continued fractions.

Finally, David Eugene Smith presents some of the mathematical ideas present in the first mathematics book published in the western hemisphere, the *Sumario Compendioso of 1556*, written by Juan Diez. Not only did this book provide an introduction to arithmetical methods, but it also discussed the solution of quadratic equations, where the methods were evidently taken from European algebras of the same century.

The Discovery of the Series Formula for π by Leibniz, Gregory and Nilakantha

RANJAN ROY

Mathematics Magazine 63 (1990), 291–306

1 Introduction

The formula for π mentioned in the title of this article is

$$\frac{\pi}{4} = 1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \cdots \quad (1)$$

One simple and well-known modern proof goes as follows:

$$\begin{aligned} \arctan x &= \int_0^x \frac{1}{1+t^2} dt \\ &= x - \frac{x^3}{3} + \frac{x^5}{5} - \cdots + (-1)^n \frac{x^{2n+1}}{2n+1} \\ &\quad + (-1)^{n+1} \int_0^x \frac{t^{2n+2}}{1+t^2} dt. \end{aligned}$$

The last integral tends to zero if $|x| \leq 1$, for

$$\begin{aligned} \left| \int_0^x \frac{t^{2n+2}}{1+t^2} dt \right| &\leq \left| \int_0^x t^{2n+2} dt \right| \\ &= \frac{|x|^{2n+3}}{2n+3} \rightarrow 0 \quad \text{as } n \rightarrow \infty. \end{aligned}$$

Thus, $\arctan x$ has an infinite series representation for $|x| \leq 1$:

$$\arctan x = x - \frac{x^3}{3} + \frac{x^5}{5} - \frac{x^7}{7} + \cdots \quad (2)$$

The series for $\pi/4$ is obtained by setting $x = 1$ in (2). The series (2) was obtained independently by Gottfried Wilhelm Leibniz (1646–1716), James Gregory (1638–1675) and an Indian mathematician of the fourteenth century or probably the fifteenth century whose identity is not definitely known. Usually ascribed to Nilakantha, the Indian proof of (2) appears to date from the mid-fifteenth century and was

a consequence of an effort to rectify the circle. The details of the circumstances and ideas leading to the discovery of the series by Leibniz and Gregory are known. It is interesting to go into these details for several reasons. The infinite series began to play a role in mathematics only in the second half of the seventeenth century. Prior to that, particular cases of the infinite geometric series were the only ones to be used. The arctan series was obtained by Leibniz and Gregory early in their study of infinite series and, in fact, before the methods and algorithms of calculus were fully developed. The history of the arctan series is, therefore, important because it reveals early ideas on series and their relationship with quadrature or the process of finding the area under a curve. In the case of Leibniz, it is possible to see how he used and transformed older ideas on quadrature to develop his methods. Leibniz's work, in fact, was primarily concerned with quadrature; the $\pi/4$ series resulted (in 1673) when he applied his method to the circle. Gregory, by comparison, was interested in finding an infinite series representation of any given function and discovered the relationship between this and the successive derivatives of the given function. Gregory's discovery, made in 1671, is none other than the Taylor series; note that Taylor was not born until 1685. The ideas of calculus, such as integration by parts, change of variables, and higher derivatives, were not completely understood in the early 1670s. Some particular cases were known, usually garbed in geometric language. For example, the fundamental theorem of calculus was stated as a geometric theorem in a work of Gregory's written in 1668. Similar examples can also be seen in a book by Isaac Barrow, Newton's mentor, published in 1670. Of course, very

soon after this transitional period, Leibniz began to create the techniques, algorithms and notations of calculus as they are now known. He had been preceded by Newton, at least as far as the techniques go, but Newton did not publish anything until considerably later. It is, therefore, possible to see how the work on arctan was at once dependent on earlier concepts and a transitional step toward later ideas.

Finally, although the proofs of (2) by Leibniz, Gregory and Nilakantha are very different in approach and motivation, they all bear a relation to the modern proof given above.

2 Gottfried Wilhelm Leibniz (1646–1716)

Leibniz’s mathematical background [1] at the time he found the $\pi/4$ formula can be quickly described. He had earned a doctor’s degree in law in February 1667, but had studied mathematics on his own. In 1672, he was a mere amateur in mathematics. That year, he visited Paris and met Christiaan Huygens (1629–1695), the foremost physicist and mathematician in continental Europe. Leibniz told the story of this meeting in a 1679 letter to the mathematician Tschirnhaus, “at that time . . . I did not know the correct definition of the center of gravity. For, when by chance I spoke of it to Huygens, I let him know that I thought that a straight line drawn through the center of gravity always cut a figure into two equal parts. . . . Huygens laughed when he heard this, and told me that nothing was further from the truth. So I, excited by this stimulus, began to apply myself to the study of the more intricate geometry, although as a matter of fact I had not at that time really studied the *Elements* [Euclid] . . . Huygens, who thought me a better geometer than I was, gave me to read the letters of Pascal, published under the name of Dettonville; and from these I gathered the method of indivisibles and centers of gravity, that is to say the well-known methods of Cavalieri and Guldinus.” [2]

The study of Pascal played an important role in Leibniz’s development as a mathematician. It was from Pascal that he learned the ideas of the “characteristic triangle” and “transmutation.” In order to understand the concept of transmutation, suppose A and B are two areas (or volumes) which have been divided up into indivisibles usually taken to be infinitesimal rectangles (or prisms). If there is a one-to-one correspondence between the indivisibles of

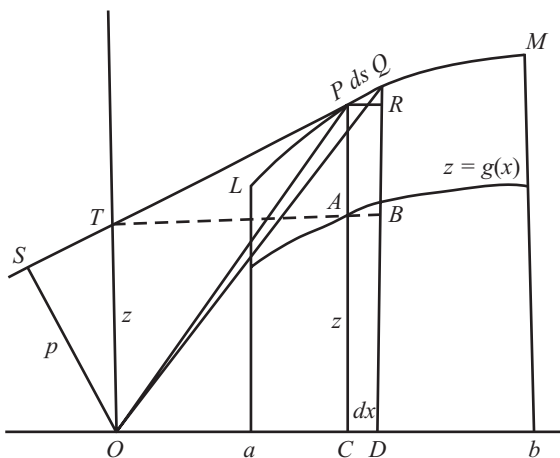


Figure 1.

A and B and if these indivisibles have equal areas (or volumes), then B is said to be obtained from A by transmutation and it follows that A and B have equal areas (or volumes). Pascal had also considered infinitesimal triangles and shown their use in finding, among other things, the area of the surface of a sphere. Leibniz was struck by the idea of an infinitesimal triangle and its possibilities. He was able to derive an interesting transmutation formula, which he then applied to the quadrature of a circle and thereby discovered the series for π . To obtain the transmutation formula, consider two neighboring points $P(x, y)$, and $Q(x + dx, y + dy)$ on a curve $y = f(x)$. First Leibniz shows that $\text{area}(\Delta OPQ) = (1/2) \text{area}(\text{rectangle}(ABCD))$. See Figure 1. Here PT is tangent to $y = f(x)$ at P and OS is perpendicular to PT . Let p denote the length of OS and z that of $AC = BD = \text{ordinate of } T$.

Since ΔOST is similar to the characteristic ΔPQR ,

$$\frac{dx}{p} = \frac{ds}{z},$$

where ds is the length of PQ . Thus,

$$\text{area}(OPQ) = \frac{1}{2}pds = \frac{1}{2}zdx. \quad (3)$$

Now, observe that for each point P on $y = f(x)$ there is a corresponding point A . Thus, as P moves from L to M , the points A form a curve, say $Z = g(x)$. If sector OLM denotes the closed region formed by $y = f(x)$ and the straight lines OL and OM , then (3) implies that

$$\text{area}(\text{sector } OLM) = \frac{1}{2} \int_a^b g(x) dx. \quad (4)$$

This is the transmutation formula of Leibniz. From (4), it follows that the area under $y = f(x)$ is

$$\int_a^b y dx = \frac{b}{2}f(b) - \frac{a}{2}f(a) + \text{area (sector } OLM) \\ = \frac{1}{2} \left([xy]_a^b + \int_a^b z dx \right). \quad (5)$$

This is none other than a particular case of the formula for integration by parts. For it is easily seen from Figure 1 that

$$z = y - x \frac{dy}{dx}. \quad (6)$$

Substituting this value of z in (5), it follows that

$$\int_a^b y dx = [xy]_a^b - \int_{f(a)}^{f(b)} x dy,$$

which is what one gets on integration by parts.

Now consider a circle of radius 1 and center (1, 0). Its equation is $y^2 = 2x - x^2$. In this case, (6) implies that

$$z = \sqrt{2x - x^2} - \frac{x(1 - x)}{\sqrt{2x - x^2}} \\ = \sqrt{\frac{x}{2 - x}} = \frac{x}{y}, \quad (7)$$

so that

$$x = \frac{2z^2}{1 + z^2}. \quad (8)$$

In Figure 2, let $\angle AOB = 2\theta$. Then the area of the sector $AOB = \theta$ and

$$\theta = \text{area}(\triangle AOB) \\ + \text{area (region between arc } AB \text{ and line } AB).$$

By the transmutation formula (4), the second area is $\frac{1}{2} \int_0^x z dt$ where z is given by (7). Now, from Figure

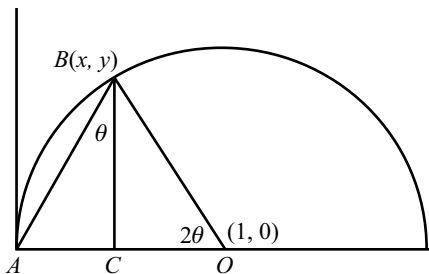


Figure 2.

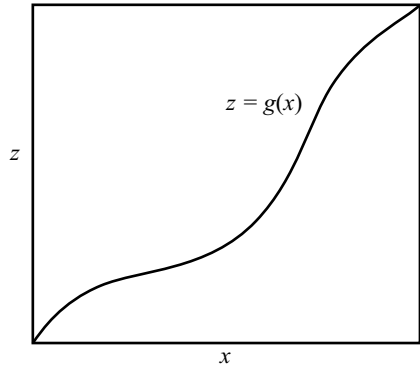


Figure 3.

3 it is seen that

$$\frac{1}{2} \int_0^x z dt = \frac{1}{2} \left(xz - \int_0^z x du \right). \quad (10)$$

Using (8) and (10), it is now possible to rewrite (9) as

$$\theta = \frac{1}{2}y + \frac{1}{2}xz - \int_0^z \frac{t^2}{1 + t^2} dt \\ = \frac{1}{2}[z(2 - x) + xz] - \int_0^z \frac{t^2}{1 + t^2} dt \\ \text{(since } y = z(2 - x)) \\ = z - \int_0^z \frac{t^2}{1 + t^2} dt.$$

At this point, Leibniz was able to use a technique employed by Nicolaus Mercator (1620–1687). The latter had considered the problem of the quadrature of the hyperbola $y(1 + x) = 1$. Since it was already known that

$$\int_0^a x^n dx = \frac{a^{n+1}}{n + 1},$$

he solved the problem by expanding $1/(1 + x)$ as an infinite series and integrating term by term. He simultaneously had the expansion for $\log(1 + x)$. Mercator published this result in 1668, though he probably had obtained it a few years earlier. A year later, John Wallis (1616–1703) determined the values of x for which the series is valid. Thus Leibniz found that

$$\theta = z - \frac{z^3}{3} + \frac{z^5}{5} - \dots \quad (11)$$

In Figure 2, $\angle ABC = \theta$ and $z = x/y = \tan \theta$. Therefore, (11) is the series for $\arctan z$.

Of course, Leibniz did not invent the notation for the integral and differential used above until 1675, and his description of the procedures is geometrical but the ideas are the same.

The discovery of the infinite series for π was Leibniz's first great achievement. He communicated his result to Huygens, who congratulated him, saying that this remarkable property of the circle will be celebrated among mathematicians forever. Even Isaac Newton (1642–1727) praised Leibniz's discovery. In a letter of October 24, 1676, to Henry Oldenburg, secretary of the Royal Society of London, he writes, "Leibniz's method for obtaining convergent series is certainly very elegant, and it would have sufficiently revealed the genius of its author, even if he had written nothing else." [3] Of course, for Leibniz this was only a first step to greater things as he himself says in his "Historia et origo calculi differentialis."

3 James Gregory (1638–1675)

Leibniz had been anticipated in the discovery of the series for \arctan by the Scottish mathematician, James Gregory, though the latter did not note the particular case for $\pi/4$ [4]. Since Gregory did not publish most of his work on infinite series and also because he died early and worked in isolation during the last seven years of his life, his work did not have the influence it deserved. Gregory's early scientific interest was in optics about which he wrote a masterly book at the age of twenty-four. His book, the *Optica Promota*, contains the earliest description of a reflecting telescope. It was in the hope, which ultimately remained unfulfilled, of constructing such an instrument that he traveled to London in 1663 and made the acquaintance of John Collins (1624–1683), an accountant and amateur mathematician. This friendship with Collins was to prove very important for Gregory when the latter was working alone at St. Andrews University in Scotland. Collins kept him abreast of the work of the English mathematicians such as Isaac Newton, John Pell (1611–1685) and others with whom Collins was in contact. [5]

Gregory spent the years 1664–1668 in Italy and came under the influence of the Italian school of geometry founded by Cavalieri. It was from Stefano degli Angeli (1623–1697), a student of Cavalieri, that Gregory learned about the work of Pierre de Fermat (1601–1665), Cavalieri, Evangelista Torricelli (1608–1647) and others. While in Italy, he wrote

two books: *Vera Circuli et Hyperbolae Quadratura* in 1667, and *Geometriae Pars Universalis* in 1668. The first book contains some highly original ideas. Gregory attempted to show that the area of a general sector of an ellipse, circle or hyperbola could not be expressed in terms of the areas of the inscribed and circumscribed triangle and quadrilateral using arithmetical operations and root extraction. The attempt failed but Gregory introduced a number of important ideas such as convergence and algebraic and transcendental functions. The second book contains the first published statement and proof of the fundamental theorem of calculus in geometrical form. It is known that Newton had discovered this result not later than 1666, although he did not make it public until later.

Gregory returned to London in the summer of 1668; Collins then informed him of the latest discoveries of mathematicians working in England, including Mercator's recently published proof of

$$\log(1+x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \dots$$

Meditation on these discoveries led Gregory to publish his next book, *Exercitationes Geometricae*, in the winter of 1668. This is a sequel to the *Pars Universalis* and is mainly about the logarithmic function and its applications. It contains, for example, the first evaluations of the indefinite integrals of $\sec x$ and $\tan x$. [6] The results are stated in geometric form.

In the autumn of 1668, Gregory was appointed to the chair in St. Andrews and he took up his duties in the winter of 1668/1669. He began regular correspondence with Collins soon after this, communicating to him his latest mathematical discoveries and requesting Collins to keep him informed of the latest activities of the English mathematicians. Thus, in a letter of March 24, 1670, Collins writes, "Mr. Newton of Cambridge sent the following series for finding the Area of a Zone of a Circle to Mr. Dary, to compare with the said Dary's approaches, putting R the radius and B the parallel distance of a Chord from the Diameter the Area of the space or Zone between them is =

$$2RB - \frac{B^3}{3R} - \frac{B^5}{20R^3} - \frac{B^7}{56R^5} - \frac{5B^9}{576R^7} \dots" [7]$$

This is all Collins writes about the series but it is, in fact, the value of the integral $2 \int_0^B (R^2 - x^2)^{1/2} dx$ after expanding by the binomial theorem and term by term integration. Newton had discovered the binomial expansion for fractional exponents in the win-

ter of 1664/1665, but it was first made public in the aforementioned letter of 1676 to Oldenburg.

There is evidence that Gregory had rediscovered the binomial theorem by 1668 [8]. However, it should be noted that the expansion for $(1 - x)^{1/2}$ does not necessarily imply a knowledge of the binomial theorem. Newton himself had proved the expansion by applying the well-known method for finding square roots of numbers to the algebraic expression $1 - x$. Moreover, it appears that the expansion of $(1 - x)^{1/2}$ was discovered by Henry Briggs (1556-1630) in the 1620's, while he was constructing the log tables [9]. But there is no indication that Gregory or Newton knew of this. In any case, for reasons unknown, Gregory was unable to make anything of the series – as evidenced by his reply of April 20, "I cannot understand the series you sent me of the circle, if this be the original, I take it to be no series." [10] However, by September 5, 1670, he had discovered the general interpolation formula, now called the Gregory-Newton interpolation formula, and had made from it a number of remarkable deductions. He now knew how "to find the sinus having the arc and to find the number having the logarithm." The latter result is precisely the binomial expansion for arbitrary exponents. For, if we take x as the logarithm of y to the base $1 + d$, then $y = (1 + d)^x$ and Gregory gives the solution as

$$(1 + d)^x = 1 + dx + \frac{x(x - 1)}{1 \cdot 2}d^2 + \frac{x(x - 1)(x - 2)}{1 \cdot 2 \cdot 3}d^3 + \dots \quad [11]$$

It is possible that Newton's series in Collins' letter had set Gregory off on the course of these discoveries, but he did not even at this point see that he could deduce Newton's result. Soon after, he did observe this and wrote on December 19, 1670, "I admire much my own dullness, that in such a considerable time I had not taken notice of this." [12] All this time, he was very eager to learn about Newton's results on series and particularly the methods he had used. Finally on December 24, 1670, Collins sent him Newton's series for $\sin x$, $\cos x$, $\sin^{-1} x$, and $x \cot x$, adding that Newton had a universal method which could be applied to any function. Gregory then made a concentrated effort to discover a general method for himself. He succeeded. In a famous letter of February 15, 1671 to Collins he writes, "As for Mr. Newton's universal method, I imagine I have some knowledge of it, both as to geometrick and mechanick curves, how-

ever I thank you for the series ye sent me and send you these following in requital." [13] Gregory then gives the series for $\arctan x$, $\tan x$, $\sec x$, $\log \sec x$, $\log \tan x$, $\sec^{-1}(\sqrt{2}e^x)$ and $2 \arctan \tanh x/2$. However, what he had found was not Newton's method but rather the Taylor expansion more than forty years before Brook Taylor (1685–1731). Newton's method consisted of reversion of series, expansion by the binomial theorem, long division by series and term by term integration [14]. Thinking that he had rediscovered Newton's method, Gregory did not publish his results. It is only from notes that he made on the back of a letter from Gedeon Shaw, an Edinburgh stationer, dated January 29, 1671, that it is possible to conclude that Gregory had the idea of the Taylor series. These notes contain the successive derivatives of $\tan x$, $\sec x$, and the other functions whose expansions he sent to Collins. The following extract from the notes gives the successive derivatives of $\tan \theta$; here m is successively y , $\frac{dy}{d\theta}$, $\frac{d^2y}{d\theta^2}$, etc., and $q = r \tan \theta$. Gregory writes [15]:

$$\begin{aligned} \text{1st: } m &= q \\ \text{2nd: } m &= r + \frac{q^2}{r} \\ \text{3rd: } m &= 2q + \frac{2q^3}{r^2} \\ \text{4th: } m &= 2r + \frac{8q^2}{r} + \frac{6q^4}{r^3} \\ \text{5th: } m &= 16q + \frac{40q^3}{r^2} + \frac{24q^5}{r^4} \\ \text{6th: } m &= 16r + \frac{136q^2}{r} + \frac{240q^4}{r^3} + \frac{120q^6}{r^5} \\ \text{7th: } m &= 272q + 987\frac{q^3}{r^2} + 1680\frac{q^5}{r^4} + 720\frac{q^7}{r^6} \\ \text{8th: } m &= 272r + 3233\frac{q^2}{r} + 11361\frac{q^4}{r^3} \\ &\quad + 13440\frac{q^6}{r^5} + 5040\frac{q^8}{r^7}. \end{aligned}$$

It is clear from the form in which the successive derivatives are written that each one is formed by multiplying the derivative with respect to q of the preceding term by $r + q^2/r$. Now writing $a = r\theta$, Gregory gives the series in the letter to Collins as follows:

$$\begin{aligned} r \tan \theta &= a + \frac{a^3}{3r^2} + \frac{2a^5}{15r^4} + \frac{17a^7}{315r^6} \\ &\quad + \frac{3233a^9}{181440r^8} + \dots \end{aligned}$$

The reasons for supposing that these notes were writ-

ten not much before he wrote to Collins and were used to construct the series are (i) the date of Gedeon Shaw's letter and (ii) Gregory's error in computing the coefficient of q^3 in the 7th m , which should be 1232 instead of 987 and which, in turn, leads to the error in the 8th m , where the coefficient of q^2/r should be 3968 instead of 3233. This error is then repeated in the series showing the origin of the series. Moreover, in the early parts of the notes, Gregory is unsure about how he should write the successive derivatives. For example, he attempts to write the derivative of $\sec \theta$ as a function of $\sec \theta$ but then abandons the idea. He comes back to it later and sees that it is easier to work with m^2 instead of m since the m^2 's can be expressed as polynomials in $\tan \theta$. This is, of course, sufficient to give him the series for $\sec \theta$. The series for $\log \sec \theta$ and $\log \tan(\pi/4 + \theta)$ he then obtains by term by term integration of the series for $\tan \theta$ and $\sec \theta$ respectively. Naturally, the 3233 error is repeated. He must have obtained the series for $\arctan x$ from the 2nd m which can be written as

$$\frac{da}{dq} = \frac{r^2}{r^2 + q^2} = 1 - \frac{q^2}{r^2} + \frac{q^4}{r^4} - \dots$$

The arctan series follows after term by term integration. Clearly, Gregory had made great progress in the study of infinite series and the calculus and, had he lived longer and published his work, he might have been classed with Newton and Leibniz as a co-discoverer of the calculus. Unfortunately, he was struck by a sudden illness, accompanied with blindness, as he was showing some students the satellites of Jupiter. He did not recover and died soon after in October, 1675, at the age of thirty-seven.

4 Kerala Gargya Nilakantha (c. 1450–c. 1550)

Another independent discovery of the series for $\arctan x$ and other trigonometric functions was made by mathematicians in South India during the fifteenth century. The series are given in Sanskrit verse in a book by Nilakantha called *Tantrasangraha* and a commentary on this work called *Tantrasangraha-vakhya* of unknown authorship. The theorems are stated without proof but a proof of the arctangent, cosine and sine series can be found in a later work called *Yuktibhasa*. This was written in Malayalam, the language spoken in Kerala, the southwest coast of India, by Jyesthadeva (c. 1500–

c. 1610) and is also a commentary on the *Tantrasangraha*. These works were first brought to the notice of the western world by an Englishman named C. M. Whish in 1835. Unfortunately, his paper on the subject had almost no impact and went unnoticed for almost a century when C. Rajagopal [16] and his associates began publishing their findings from a study of these manuscripts. The contributions of medieval Indian mathematicians are now beginning to be recognized and discussed by authorities in the field of the history of mathematics [17].

It appears from the astronomical data contained in the *Tantrasangraha* that it was composed around the year 1500. The *Yuktibhasa* was written about a century later. It is not completely clear who the discoverer of these series was. In the *Aryabhatiya-bhasya*, a work on astronomy, Nilakantha attributes the series for sine to Madhava. This mathematician lived between the years 1340–1425. It is not known whether Madhava found the other series as well or whether they are somewhat later discoveries.

Little is known about these mathematicians. Madhava lived near Cochin in the very southern part of India (Kerala) and some of his astronomical work still survives. Nilakantha was a versatile genius who wrote not only on astronomy and mathematics but also on philosophy and grammar. His erudite expositions on the latter subjects were well known and studied until recently. He attracted several gifted students, including Tuncath Ramantijan Ezuthassan, an early and important figure in Kerala literature. About Jyesthadeva, nothing is known except that he was a Brahmin of the house of Parakroda.

In the *Tantrasangraha-vakhya*, the series for arctan, sine and cosine are given in verse which, when converted to mathematical symbols may be written as follows (see Figure 4):

$$r \arctan \frac{y}{x} = \frac{1}{1} \cdot \frac{ry}{x} - \frac{1}{3} \cdot \frac{ry^3}{x^3} + \frac{1}{5} \cdot \frac{ry^5}{x^5} - \dots, \quad \text{where } \frac{y}{x} \leq 1,$$

$$y = s - s \cdot \frac{s^2}{(2^2 + 2)r^2} + s \cdot \frac{s^2}{(2^2 + 2)r^2} \cdot \frac{s^2}{(4^2 + 4)r^2} - \dots \quad (\text{sine})$$

$$r - x = r \cdot \frac{s^2}{(2^2 - 2)r^2} - r \cdot \frac{s^2}{(2^2 - 2)r^2} \cdot \frac{s^2}{(4^2 - 4)r^2} + \dots \quad (\text{cosine})$$

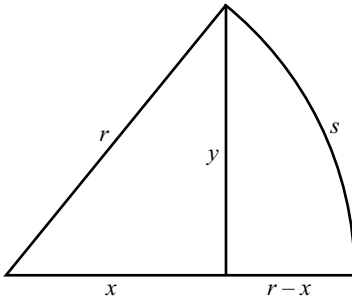


Figure 4.

There are also some special features in the *Tantrasangraha*'s treatment of the series for $\pi/4$ which were not considered by Leibniz and Gregory. Nilakantha states some rational approximations for the error incurred on taking only the first n terms of the series. The expression for the approximation is then used to transform the series for $\pi/4$ into one which converges more rapidly. The errors are given as follows:

$$\frac{\pi}{4} \approx 1 - \frac{1}{3} + \frac{1}{5} - \dots \mp \frac{1}{n} \pm f_i(n+1) \quad i = 1, 2, 3, \quad (12)$$

where

$$f_1(n) = \frac{1}{2n}, \quad f_2(n) = \frac{n/2}{n^2 + 1}, \quad \text{and}$$

$$f_3(n) = \frac{(n/2)^2 + 1}{(n^2 + 5)n/2}.$$

The transformed series are as follows:

$$\frac{\pi}{4} = \frac{3}{4} + \frac{1}{3^2 - 3} - \frac{1}{5^3 - 5} + \frac{1}{7^3 - 7} - \dots \quad (13)$$

and

$$\frac{\pi}{4} = \frac{4}{1^5 + 4 \cdot 1} - \frac{4}{3^5 + 4 \cdot 3} + \frac{4}{5^5 + 4 \cdot 5} - \dots$$

Leibniz's proof of the formula for $\pi/4$ was found by the quadrature of a circle. The proof in Jyesthadeva's book is by a direct rectification of an arc of a circle. In Figure 5, the arc AC is a quarter circle of radius one with center O and $OABC$ is a square. The side AB is divided into n equal parts of length δ so that $n\delta = 1, P_{r-1}P_r = \delta$. EF and $P_{r-1}D$ are perpendicular to OP_r . Now, the triangles OEF and $OP_{r-1}D$ are similar, which gives

$$\frac{ER}{OE} = \frac{P_{r-1}D}{OP_{r-1}}, \quad \text{that is} \quad EF = \frac{P_{r-1}D}{OP_{r-1}}.$$

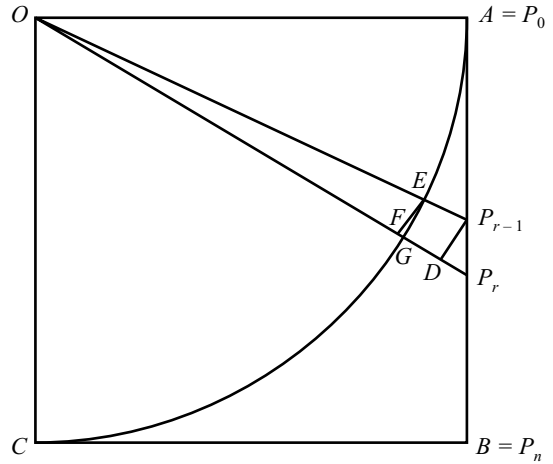


Figure 5.

The similarity of the triangles $P_{r-1}P_rD$ and OAP_r gives

$$\frac{P_{r-1}P_r}{OP_r} = \frac{P_{r-1}D}{OA} \quad \text{or} \quad P_{r-1}D = \frac{P_{r-1}P_r}{OP_r}.$$

Thus

$$EF = \frac{P_{r-1}P_r}{OP_{r-1}OP_r} \approx \frac{P_{r-1}P_r}{OP_r^2} = \frac{\delta}{1 + AP_r^2} = \frac{\delta}{1 + r^2\delta^2}.$$

Since arc $EG \approx EF \approx \delta/(1+r^2\delta^2)$, $\frac{1}{8}$ arc of circle is

$$\pi/4 = \lim_{n \rightarrow \infty} \sum_{r=1}^n \frac{\delta}{1 + r^2\delta^2}. \quad (14)$$

Of course, a clear idea of limits did not exist at that time so that the relation was understood in an intuitive sense only. To evaluate the limit, Jyesthadeva uses two lemmas. One is the geometric series

$$\frac{1}{1+x} = 1 - x + x^2 - x^3 + \dots$$

Jyesthadeva says that the expansion is obtained on iterating the following procedure:

$$\begin{aligned} \frac{1}{1+x} &= 1 - x \left(\frac{1}{1+x} \right) \\ &= 1 - x \left(1 - x \left(\frac{1}{1+x} \right) \right). \end{aligned}$$

The other result is that

$$\begin{aligned} S_n^{(p)} &\equiv 1^p + 2^p + \dots + n^p \\ &\sim \frac{n^{p+1}}{p+1} \quad \text{for large } n. \end{aligned} \quad (15)$$

A sketch of a proof is given by Jyesthadeva. He notes first that

$${}_nS_n^{(p-1)} = S_n^{(p)} + S_1^{(p-1)} + S_2^{(p-1)} + \dots + S_{n-1}^{(p-1)}. \tag{16}$$

This is easy to verify. Relation (16) is also contained in the work of the tenth century Arab mathematician Alhazen, who gives a geometrical proof in the Greek tradition [18]. He uses it to evaluate $S_n^{(3)}$ and $S_n^{(4)}$ which occur in a problem about the volume of certain solid of revolution. *Yuktibhasa* shows that for $p = 2, 3$

$$S_1^{(p-1)} + S_2^{(p-1)} + \dots + S_{n-1}^{(p-1)} \sim \frac{S_n^{(p)}}{p}, \tag{17}$$

and then suggests that by induction the result will be true for all values of p . Once this is granted, it follows that if

$$S_n^{(p-1)} \sim \frac{n^p}{p},$$

then by (16) and (17),

$${}_nS_n^{(p-1)} \sim S_n^{(p)} + \frac{S_n^{(p)}}{p} \text{ or } S_n^{(p)} \sim \frac{n^{p+1}}{p+1},$$

and (15) is inductively proved.

We now note that (14) can be rewritten, after expanding $1/(1+r^2\delta^2)$ into a geometric series, as

$$\begin{aligned} \frac{\pi}{4} &= \lim_{n \rightarrow \infty} \left[\delta \sum_{r=1}^n 1 - \delta^3 \sum_{r=1}^n r^2 + \delta^5 \sum_{r=1}^n r^4 - \dots \right] \\ &= \lim_{n \rightarrow \infty} \left[1 - \frac{1}{n^3} \sum_{r=1}^n r^2 + \frac{1}{n^5} \sum_{r=1}^n r^4 - \dots \right] \\ &= 1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \dots, \end{aligned}$$

where we have used relation (15) and the fact that $\delta = 1/n$. Now consider the approximation (12) and its application to the transformation of series. Suppose that

$$\sigma_n = 1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \dots \pm \frac{1}{n} \mp f(n+1),$$

where $f(n+1)$ is a rational function of n which will make σ_n a better approximation of $\pi/4$ than the n th partial sum S_n . Changing n to $n-2$ we get

$$\sigma_{n-2} = 1 - \frac{1}{3} + \frac{1}{5} - \dots \mp \frac{1}{n-2} \pm f(n-1).$$

Subtracting the second relation from the first,

$$\pm u_n = \sigma_n - \sigma_{n-2} = \pm \frac{1}{n} \mp f(n+1) \mp f(n-1). \tag{18}$$

Then

$$\begin{aligned} \sigma_n &= \sigma_{n-2} \pm u_n \\ &= \sigma_{n-4} \mp u_{n-2} \pm u_n \\ &= \dots \\ &= \sigma_1 - u_3 + u_5 - u_7 + \dots \pm u_n \\ &= 1 - f(2) - u_3 + u_5 - u_7 + \dots \pm u_n. \end{aligned}$$

It is clear that

$$\lim_{n \rightarrow \infty} \sigma_n = \frac{\pi}{4}$$

and therefore

$$\frac{\pi}{4} = 1 - f(2) - u_3 + u_5 - u_7 + \dots. \tag{19}$$

Thus, we have a new series for $\pi/4$ which depends on how the function $f(n)$ is chosen. Naturally, the aim is to choose $f(n)$ in such a way that (19) is more rapidly convergent than (1). This is the idea behind the series (13). Now equation (18) implies that

$$f(n+1) + f(n-1) = \frac{1}{n} - u_n. \tag{20}$$

For (19) to be more rapidly convergent than (1), u_n should be $o(1/n)$, that is negligible compared to $1/n$. It is reasonable to assume $f(n+1) \approx f(n-1) \approx f(n)$. These observations together with (20) imply that $f(n) = 1/2n$ is a possible rational approximation in equation (12). With this $f(n)$, the value of u_n is given by (20) to be

$$u_n = \frac{1}{n} - \frac{1}{2(n+1)} - \frac{1}{2(n-1)} = -\frac{1}{n^3 - n}.$$

Substituting this in (19) gives us (13), which is

$$\frac{\pi}{4} = 1 - \frac{1}{4} + \frac{1}{3^3 - 3} - \frac{1}{5^3 - 5} + \frac{1}{7^3 - 7} - \dots.$$

The other series

$$\frac{\pi}{4} = \frac{4}{1^5 + 4 \cdot 1} - \frac{4}{3^5 + 4 \cdot 3} + \frac{4}{5^5 + 4 \cdot 5} - \dots$$

is obtained by taking $f(n) = (n/2)/(n^2 + 1)$ in (19).

It should be mentioned that Newton was aware of the correction $f_1(n) = 1/2n$. For in the letter to Oldenburg, referred to earlier, he says, "By the series of Leibniz also if half the term in the last place be

added and some other like devices be employed, the computation can be carried to many figures.” However, he says nothing about transforming the series by means of this correction.

It appears that Nilakantha was aware of the impossibility of finding a finite series of rational numbers to represent π . In the *Aryabhatiya-bhasya* he writes, “If the diameter, measured using some unit of measure, were commensurable with that unit, then the circumference would not likewise allow itself to be measured by means of the same unit; so likewise in the case where the circumference is measurable by some unit, then the diameter cannot be measured using the same unit.” [19]

The *Yuktibhasa* contains a proof of the arctan series also and it is obtained in exactly the same way except that one rectifies only a part of the $1/8$ circle.

It can be shown that if $\pi/4 = S_n + f(n)$, where S_n is the n th partial sum, then $f(n)$ has the continued fraction representation

$$f(n) = \frac{1}{2} \left[\frac{1}{n+} \frac{1^2}{n+} \frac{2^2}{n+} \frac{3^2}{n+} \dots \right]. \quad (21)$$

Moreover, the first three convergents are

$$f_1(n) = \frac{1}{2n}, \quad f_2(n) = \frac{n/2}{n^2 + 1}, \quad \text{and}$$

$$f_3(n) = \frac{(n/2)^2 + 1}{(n^2 + 5)n/2},$$

which are the values quoted in (13). Clearly, Nilakantha was using some procedure which gave the successive convergents of the continued fraction (21) but the text contains no suggestion that (20) was actually known to him. This continued fraction implies that

$$\frac{2}{4 - \pi} = 2 + \frac{1^2}{2+} \frac{3^2}{2+} \frac{5^2}{2+} \dots,$$

which may be compared with the continued fraction of the seventeenth-century English mathematician, William Brouncker (1620-1684), who gave the result

$$\frac{4}{\pi} = 1 + \frac{1^2}{2+} \frac{3^2}{2+} \frac{5^2}{2+} \dots.$$

The third approximation

$$f_3(n) = \frac{(n/2)^2 + 1}{(n^2 + 5)n/2}$$

is very effective in obtaining good numerical values for π without much calculation. For example

$$1 - \frac{1}{3} + \dots - \frac{1}{19} + f_3(20)$$

gives the value of π correct up to eight decimal places [20]. Nilakantha himself gives 104348/33215 which is correct up to nine places. It is interesting that the Arab mathematician Jamshid-al-Kasi, who also lived in the fifteenth century, had obtained the same approximation by a different method.

5 Independence of these discoveries

The question naturally arises of the possibility of mutual influence between or among the discoverers of power series, in particular the series for the trigonometric functions. Because of the lively trade relations between the Arabs and the west coast of India over the centuries, it is generally accepted that mathematical ideas were also exchanged. However, there is no evidence in any existing mathematical works of the Arabs that they were aware of the concept of a power series. Therefore, we may grant the Indians priority in the discovery of the series for sine, cosine and arc tangent. Moreover, historians of mathematics are in agreement that the European mathematicians were unaware of the Indian discovery of infinite series [21]. Thus, we may conclude that Newton, Gregory and Leibniz made their discoveries independently of the Indian work. In fact, it appears that yet another independent discovery of an infinite series giving the value of π was made by the Japanese mathematician Takebe Kenko (1664–1739) in 1722. His series is

$$\pi^2 = 4 \left[1 + \sum_{n=1}^{\infty} \frac{2^{2n+1}(n!)^2}{(2n+1)!} \right]. \quad [22]$$

This series was not obtained from the arctan series and its discussion is therefore not included. However, the independent discovery of the infinite series by different persons living in different environments and cultures gives us insight into the character of mathematics as a universal discipline.

References

1. For further information about Leibniz’s mathematical development, the reader may consult: J. E. Hofmann, *Leibniz in Paris 1672–1676* (Cambridge: The Cambridge University Press, 1974) and its review by A. Weil, *Collected Papers Vol. 3* (New York: Springer-Verlag, 1979). An English translation of Leibniz’s own account, *Historia et origo calculi differentialis*, can be found in J. M. Child, *The Early Mathematical*

Manuscripts of Leibniz (Chicago: Open Court, 1920). An easily available synopsis of Leibniz's work in calculus is given in C. H. Edwards, Jr., *The Historical Development of the Calculus* (New York: Springer-Verlag, 1979).

2. *The Early Mathematical Manuscripts*, p. 215. Bonaventura Cavalieri (1598–1647) published his *Geometrica Indivisibilibus* in 1635. This book was very influential in the development of calculus. Cavalieri's work indicated that

$$\int_0^a x^n dx = \frac{a^{n+1}}{n+1},$$

when n is a positive integer.

Blaise Pascal (1623–1662) made important and fundamental contributions to projective geometry, probability theory and the development of calculus. The work to which Leibniz refers was published in 1658 and contains the first statement and proof of

$$\int_{\theta_0}^{\theta} \sin \phi d\phi = \cos \theta_0 - \cos \theta.$$

This proof is presented in D. J. Struik's *A Source Book in Mathematics 1200–1800* (Cambridge: Harvard University Press, 1969), p. 239.

Paul Guldin (1577–1643), a Swiss mathematician of considerable note, contributed to the development of calculus and his methods were generally more rigorous than those of Cavalieri.

3. See H. W. Turnbull (ed.), *The Correspondence of Isaac Newton* (Cambridge: The University Press 1960), Vol. 2, p. 130.
4. Peter Beckmann has persuasively argued that Gregory must have known the series for $\pi/4$ as well. See Beckmann's *A History of Pi* (Boulder, Colorado: The Golem Press, 1977), p. 133.
5. The reader might find it of interest to consult: H. W. Turnbull (ed.), *James Gregory Tercentenary Memorial Volume* (London: G. Bell, 1939). This volume contains Gregory's scientific correspondence with John Collins and a discussion of the former's life and work.
6. A proof of the formula

$$\int_0^{\theta} \sec \phi d\phi = \log \tan \left(\frac{\pi}{4} + \frac{\theta}{2} \right)$$

was of considerable significance and interest to mathematicians in the 1660's due to its connection with a problem in navigation. Gerhard Mercator (1512–1594) published his engraved "Great World Map" in 1569. The construction of the map employed the famous Mercator projection. Edward Wright, a Cambridge professor of mathematics, noted that the ordinate on the Mercator map corresponding to a latitude

of θ° on the globe is given by $c \int_0^{\theta} \sec \phi d\phi$, where c is suitably chosen according to the size of the map. In 1599, Wright published this result in his *Certain Errors in Navigation Corrected*, which also contained a table of latitudes computed by the continued addition of the secants of $1', 2', 3'$, etc. This approximation to $\int_0^{\theta} \sec \phi d\phi$ was sufficiently exact for the mariner's use. In the early 1640's, Henry Bond observed that the values in Wright's table could be obtained by taking the logarithm of $\tan(\pi/4 + \theta/2)$. This observation was published in 1645 in Richard Norwood's *Epitome of Navigation*. A theoretical proof of this observation was very desirable and Nicolaus Mercator had offered a sum of money for its demonstration in 1666. John Collins, who had himself written a book on navigation, drew Gregory's attention to this problem and, as we noted, Gregory supplied a proof. For more details, one may consult the following two articles by F. Cajori: On an Integration antedating the Integral Calculus, *Bibliotheca Mathematica* 14 (1913/14), 312–19, and Algebra In Napier's Day and Alleged Prior Invention of Logarithms, in C. G. Knott (ed.), *Napier Memorial Volume* (London: Longmans, Green & Co., 1915), pp. 93–106. More recently, J. Lohne has established that Thomas Harriot (1560–1621) had evaluated the integral $\int_0^{\theta} \sec \phi d\phi$ in 1594 by a stereographic projection of a spherical loxodrome from the south pole into a logarithmic spiral. This work was unpublished and remained unknown until Lohne brought it to light. See J. A. Lohne, Thomas Harriot als Mathematiker, *Centaurus*, 11 (1965–66), 19–45. Thus it happened that, although $\int \sec \phi d\phi$ is a relatively difficult trigonometric integral, it was the first one to be discovered.

7. *James Gregory*, p. 89.
8. See *The Correspondence of Isaac Newton*, Vol. 1, p. 52, note 1.
9. See D. T. Whiteside, Henry Briggs: The Binomial theorem Anticipated, *The Mathematical Gazette* 15 (1962), 9. Whiteside shows how the expansion of $(1+x)^{1/2}$ arose out of Briggs' work on logarithms.
10. *James Gregory*, p. 92.
11. In their review of the Gregory Memorial Volume, M. Dehn and E. Hellinger explain how the binomial expansion comes out of the interpolation formula. See *The American Mathematical Monthly* 50 (1943), 149.
12. *James Gregory*, p. 148.
13. *Ibid.*, p. 170.
14. It should be mentioned that Newton himself discovered the Taylor series around 1691. See D. T. Whiteside (ed.), *The Mathematical Papers of Isaac Newton*, Vol. VII (Cambridge: The Cambridge University Press, 1976), p. 19. In fact, Taylor was anticipated by at least five mathematicians. However, the Taylor

- series is not unjustly named after Brook Taylor who published it in 1715. He saw the importance of the result and derived several interesting consequences. For a discussion of these matters see L. Feigenbaum, Brook Taylor and the Method of Increments, *Archive for History of Exact Sciences*, 34 (1985), 1–140.
15. James Gregory, p. 352.
 16. Rajagopal's work may be found in the following papers: (with M. S. Rangachari) On an Untapped Source of Medieval Keralese Mathematics, *Archive for History of Exact Sciences*, 18 (1977), 89–102, On Medieval Kerala Mathematics, *Archive for History of Exact Sciences*, 35 (1986), 91–99. These papers give the Sanskrit verses of the *Tantrasangrahavakhya* which describe the series for the arctan, sine and cosine. An English translation and commentary is also provided. A commentary on the proof of arctan series given in the *Yuktibhasa* is available in the two papers: A Neglected Chapter of Hindu Mathematics, *Scripta Mathematica*, 15 (1949), 201–209; On the Hindu Proof of Gregory's Series, *Ibid.*, 17 (1951), 65–74. A commentary on the *Yuktibhasa*'s proof of the sine and cosine series is contained in C. Rajagopal and A. Venkataraman, The sine and cosine power series in Hindu mathematics, *Journal of the Royal Asiatic Society of Bengal, Science*, 15 (1949), 1–13.
 17. See J. E. Hofmann, Über eine alt indische Berechnung von π und ihre allgemeine Bedeutung, *Mathematische-Physikalische Semester Berichte*, Bd. 3, H. 3/4, Hamburg (1953). See also D. T. Whiteside, Patterns of Mathematical Thought in the later Seventeenth Century, *Archive for History of Exact Sciences*, 1 (1960–1962), 179–388. For a discussion of medieval Indian mathematicians and the *Tantrasangraha* in particular, one might consult: A. P. Jushkevich, *Geschichte der Mathematik in Mittelalter* (German translation Leipzig, 1964, of the Russian original, Moscow, 1961.)
 18. See *The Historical Development of the Calculus* (mentioned in note 1), p. 84. Alhazen is the latinized form of the name Ibn Al-Haytham (c. 965–1039).
 19. See *Geschichte der Mathematik*, p. 169.
 20. These observations concerning the continued fraction expansion of $f(n)$ and its relation to the Indian work and that of Brouncker, and concerning the decimal places in $f(20)$, are due to D. T. Whiteside. See On Medieval Kerala Mathematics, of note 16.
 21. See Patterns of Mathematical Thought in the later Seventeenth Century, of note 17. See also A. Weil, History of Mathematics: Why and How, in *Collected Papers*, Vol. 3 (New York: Springer-Verlag 1979), p. 435.
 22. See D. E. Smith and Y. Mikami, *A History of Japanese Mathematics* (Chicago: Open Court, 1914). This series was also obtained by the French missionary Pierre Jartoux (1670–1720) in 1720. He worked in China and was in correspondence with Leibniz, but the present opinion is that Takebe's discovery was independent. Leonhard Euler (1707–1783) rediscovered the same series in 1737. A simple evaluation of it can be given using Clausen's formula for the square of a hypergeometric series.

Ideas of Calculus in Islam and India

VICTOR J. KATZ

Mathematics Magazine 68 (1995), 163–174

1 Introduction

Isaac Newton created his version of the calculus during the years from about 1665 to 1670. One of Newton's central ideas was that of a power series, an idea he believed he had invented out of the analogy with the infinite decimal expansions of arithmetic [9, Vol. III, p. 33]. Newton, of course, was aware of earlier work done in solving the area problem, one of the central ideas of what was to be the calculus, and he knew well that the area under the curve $y = x^n$ between $x = 0$ and $x = b$ was given by $b^{n+1}/(n + 1)$. (This rule had been developed by several mathematicians in the 1630s, including Bonaventura Cavalieri, Gilles Persone de Roberval, and Pierre de Fermat.) By developing power series to represent various functions, Newton was able to use this basic rule to find the areas under a wide variety of curves. Conversely, the use of the area formula enabled him to develop power series. For example, Newton developed the power series for $y = \arcsin x$, in effect by defining it in terms of an area and using the area formula. He then produced the power series for the sine by solving the equation $y = \arcsin x$ for $x = \sin y$ by inversion of the series. What Newton did not know, however, was that both the area formula — which he believed had been developed some 35 years earlier — and the power series for the sine had been *known for hundreds of years* elsewhere in the world. In particular, the area formula had been developed in Egypt around the year A.D. 1000 and the power series for the sine, as well as for the cosine and the arctangent, had been developed in India, probably in the fourteenth century. It is the development of these two ideas that will be discussed in this article.

Before going back to eleventh-century Egypt, however, we will first review the argument used both

by Fermat and Roberval in working out their version of the area formula in 1636. In a letter to Fermat in October of that year, Roberval wrote that he had been able to find the area under curves of the form $y = x^k$ by using a formula — whose history in the Islamic world we will trace — for the sums of powers of the natural numbers: “The sum of the square numbers is always greater than the third part of the cube which has for its root the root of the greatest square, and the same sum of the squares with the greatest square removed is less than the third part of the same cube; the sum of the cubes is greater than the fourth part of the fourth power and with the greatest cube removed, less than the fourth part, etc.” [5, p. 221]. In other words, finding the area of the desired region depends on the formula

$$\sum_{i=1}^{n-1} i^k < \frac{n^{k+1}}{k+1} < \sum_{i=1}^n i^k.$$

Fermat wrote back that he already knew this result and, like Roberval, had used it to determine the area under the graph of $y = x^k$ over the interval $[0, x_0]$. Both men saw that if the base interval was divided into n equal subintervals, each of length x_0/n , and if over each subinterval a rectangle whose height is the y -coordinate of the right endpoint was erected (see Figure 1), then the sum of the areas of these circumscribed rectangles is

$$\begin{aligned} \frac{x_0^k}{n^k} \frac{x_0}{n} + \frac{(2x_0)^k}{n^k} \frac{x_0}{n} + \cdots + \frac{(nx_0)^k}{n^k} \frac{x_0}{n} \\ = \frac{x_0^{k+1}}{n^{k+1}} (1^k + 2^k + \cdots + n^k). \end{aligned}$$

Similarly, they could calculate the sum of the areas of the inscribed rectangles, those whose height is the y -coordinate of the left endpoint of the corresponding subinterval. In fact, if A is the area under

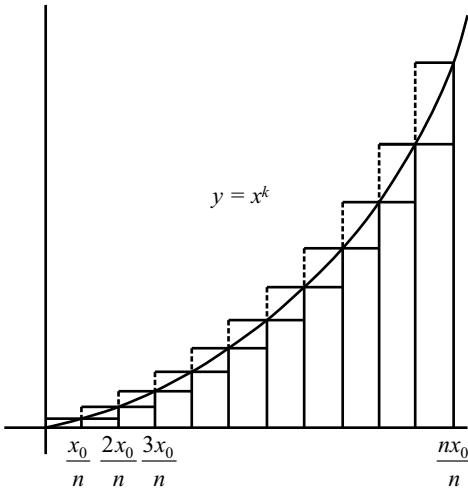


Figure 1.

the curve between 0 and x_0 , then

$$\begin{aligned} \frac{x_0^{k+1}}{n^{k+1}}(1^k + 2^k + \dots + (n-1)^k) &< A \\ &< \frac{x_0^{k+1}}{n^{k+1}}(1^k + 2^k + \dots + n^k). \end{aligned}$$

The difference between the outer expressions of this inequality is simply the area of the rightmost circumscribed rectangle. Because x_0 and $y_0 = x_0^k$ are fixed, Fermat knew that the difference could be made less than any assigned value simply by taking n sufficiently large. It follows from the inequality cited by Roberval that both the area A and the value $x_0^{k+1}/(k+1) = x_0 y_0 / (k+1)$ are squeezed between two values whose difference approaches 0. Thus Fermat and Roberval found that the desired area was $x_0 y_0 / (k+1)$.

The obvious question is how either of these two men discovered formulas for sums of powers. But at present, there is no answer to this question. There is nothing extant on this formula in the works of Roberval other than the letter cited, and all we have from Fermat on this topic, in letters to Marin Mersenne and Roberval, is a general statement in terms of triangular numbers, pyramidal numbers, and the other numbers that occur as columns of Pascal’s triangle. (We note that Fermat’s work was done some twenty years before Pascal published his material on the arithmetical triangle; the triangle had, however, been published in many versions in China, the Middle East, North Africa, and Europe over the previous 600 years. See [4], pp. 191–192; 241–242, 324–325.) Here is what Fermat says: “The last side

multiplied by the next greater makes twice the triangle. The last side multiplied by the triangle of the next greater side makes three times the pyramid. The last side multiplied by the pyramid of the next greater side makes four times the triangulotriangle. And so on by the same progression in infinitum” [5, p. 230]. Fermat’s statement can be written using the modern notation for binomial coefficients as

$$n \binom{n+k}{k} = (k+1) \binom{n+k}{k+1}.$$

We can derive from this formula for each k in turn, beginning with $k = 1$, an explicit formula for the sum of the k th powers by using the properties of the Pascal triangle. For example, if $k = 2$, we have

$$\begin{aligned} n \binom{n+2}{2} &= 3 \binom{n+2}{3} \\ &= 3 \sum_{j=2}^{n+1} \binom{j}{2} = 3 \sum_{j=2}^{n+1} \frac{j(j-1)}{2} \\ &= 3 \sum_{i=1}^n \frac{i(i+1)}{2} = 3 \sum_{i=1}^n \frac{i^2 + i}{2}. \end{aligned}$$

Therefore,

$$2 \frac{n}{3} \frac{(n+2)(n+1)}{2} - \sum_{i=1}^n i = \sum_{i=1}^n i^2$$

and

$$\begin{aligned} \sum_{i=1}^n i^2 &= \frac{n^3 + 3n^2 + 2n}{3} - \frac{n^2 + n}{2} \\ &= \frac{2n^3 + 3n^2 + n}{6} = \frac{n^3}{3} + \frac{n^2}{2} + \frac{n}{6}. \end{aligned}$$

In general, the sum formula is of the form

$$\sum_{i=1}^n i^k = \frac{n^{k+1}}{k+1} + \frac{n^k}{2} + p(n),$$

where $p(n)$ is a polynomial in n of degree less than k , and Roberval’s inequality can be proved for each k . We do not know if Fermat’s derivation was like that above, however, because he only states a sum formula explicitly for the case $k = 4$ and gives no other indication of his procedure.

2 Sums of integer powers in eleventh-century Egypt

The formulas for the sums of the k th powers, however, at least through $k = 4$, as well as a version

of Roberval's inequality, were developed some 650 years before the mid-seventeenth century by Abu Ali al-Hasan ibn al-Hasan ibn al-Haytham (965–1039), known in Europe as Alhazen. The formulas for the sums of the squares and cubes were stated even earlier. The one for squares was stated by Archimedes around 250 B.C. in connection with his quadrature of the parabola, while the one for cubes, although it was probably known to the Greeks, was first explicitly written down by Aryabhata in India around 500 [2, pp. 37–38]. The formula for the squares is not difficult to discover, and the one for cubes is virtually obvious, given some experimentation. By contrast, the formula for the sum of the fourth powers is not obvious. If one can discover a method for determining this formula, one can discover a method for determining the formula for the sum of any integral powers. Ibn al-Haytham showed in fact how to develop the formula for the k th powers from $k = 1$ to $k = 4$; all his proofs were similar in nature and easily generalizable to the discovery and proof of formulas for the sum of any given powers of the integers. That he did not state any such generalization is probably due to his needing only the formulas for the second and fourth powers to solve the problem in which he was interested: computing the volume of a certain paraboloid.

Before discussing ibn al-Haytham's work, it is good to briefly describe the world of Islamic science. (See [1] for more details.) During the ninth century, the Caliph al-Ma'mun established a research institute, the House of Wisdom, in Baghdad and invited scholars from all parts of the caliphate to participate in the development of a scientific tradition in Islam. These scientists included not only Moslem Arabs, but also Christians, Jews, and Zoroastrians, among others. Their goals were, first, to translate into Arabic the best mathematical and scientific works from Greece and India, and, second, by building on this base, to create new mathematical and scientific ideas. Although the House of Wisdom disappeared after about two centuries, many of the rulers of the Islamic states continued to support scientists in their quest for knowledge, because they felt that the research would be of value in practical applications.

Thus it was that ibn al-Haytham, born in Basra, now in Iraq, was called to Egypt by the Caliph al-Hakim to work on a Nile control project. Although the project never came to fruition, ibn al-Haytham did produce in Egypt his most important scientific work, the *Optics* in seven books. The *Optics* was

translated into Latin in the early thirteenth century and was studied and commented on in Europe for several centuries thereafter. Ibn al-Haytham's fame as a mathematician from the medieval period to the present chiefly rests on his treatment of "Alhazen's problem," the problem of finding the point or points on some reflecting surface at which the light from one of two points outside that surface is reflected to the other. In the fifth book of the *Optics* he set out his solutions to this problem for a variety of surfaces, spherical, cylindrical, and conical, concave and convex. His results, based on six separately proved lemmas on geometrical constructions, show that he was in full command of both the elementary and advanced geometry of the Greeks.

The central idea in ibn al-Haytham's proof of the sum formulas was the derivation of the equation

$$(n + 1) \sum_{i=1}^n i^k = \sum_{i=1}^n i^{k+1} + \sum_{p=1}^n \left(\sum_{i=1}^p i^k \right). \quad (*)$$

Naturally, he did not state this result in general form. He only stated it for particular integers, namely $n = 4$ and $k = 1, 2, 3$, but his proof for each of those k is by induction on n and is immediately generalizable to any value of k . (See [7] for details.) We consider his proof for $k = 3$ and $n = 4$:

$$\begin{aligned} & (4 + 1)(1^3 + 2^3 + 3^3 + 4^3) \\ &= 4(1^3 + 2^3 + 3^3 + 4^3) + 1^3 + 2^3 + 3^3 + 4^3 \\ &= 4 \cdot 4^3 + 4(1^3 + 2^3 + 3^3) + 1^3 + 2^3 + 3^3 + 4^3 \\ &= 4^4 + (3 + 1)(1^3 + 2^3 + 3^3) + 1^3 + 2^3 + 3^3 + 4^3. \end{aligned}$$

Because equation (*) is assumed true for $n = 3$,

$$\begin{aligned} & (3 + 1)(1^3 + 2^3 + 3^3) \\ &= 1^4 + 2^4 + 3^4 + (1^3 + 2^3 + 3^3) + (1^3 + 2^3) + 1^3. \end{aligned}$$

Equation (*) is therefore proved for $n = 4$. One can easily formulate ibn al-Haytham's argument in modern terminology to give a proof for any k by induction on n .

Ibn al-Haytham now uses his result to derive formulas for the sums of integral powers. First, he proves the sum formulas for squares and cubes:

$$\begin{aligned} \sum_{i=1}^n i^2 &= \left(\frac{n}{3} + \frac{1}{3} \right) n \left(n + \frac{1}{2} \right) \\ &= \frac{n^3}{3} + \frac{n^2}{2} + \frac{n}{6} \end{aligned}$$

$$\begin{aligned} \sum_{i=1}^n i^3 &= \left(\frac{n}{4} + \frac{1}{4}\right) n(n+1)n \\ &= \frac{n^4}{4} + \frac{n^3}{2} + \frac{n^2}{4}. \end{aligned}$$

We will not deal with these proofs here, but only with the derivation of the analogous result for the fourth powers. Although ibn al-Haytham himself derives this result only for $n = 4$, he asserts it for arbitrary n . We will therefore use modern techniques modeled on ibn al-Haytham’s method to derive it for that case. We begin by using the formulas for the sums of squares and cubes to rewrite equation (*) in the form

$$\begin{aligned} (n+1) \sum_{i=1}^n i^3 &= \sum_{i=1}^n i^4 + \sum_{p=1}^n \left(\frac{p^4}{4} + \frac{p^3}{2} + \frac{p^2}{4}\right) \\ &= \sum_{i=1}^n i^4 + \frac{1}{4} \sum_{i=1}^n i^4 \\ &\quad + \frac{1}{2} \sum_{i=1}^n i^3 + \frac{1}{4} \sum_{i=1}^n i^2. \end{aligned}$$

It then follows that

$$\begin{aligned} (n+1) \sum_{i=1}^n i^3 &= \frac{5}{4} \sum_{i=1}^n i^4 + \frac{1}{2} \sum_{i=1}^n i^3 + \frac{1}{4} \sum_{i=1}^n i^2 \\ \frac{5}{4} \sum_{i=1}^n i^4 &= \left(n+1 - \frac{1}{2}\right) \sum_{i=1}^n i^3 - \frac{1}{4} \sum_{i=1}^n i^2 \\ \sum_{i=1}^n i^4 &= \frac{4}{5} \left(n + \frac{1}{2}\right) \sum_{i=1}^n i^3 - \frac{1}{5} \sum_{i=1}^n i^2 \\ &= \frac{4}{5} \left(n + \frac{1}{2}\right) \left(\frac{n}{4} + \frac{1}{4}\right) n(n+1)n \\ &\quad - \frac{1}{5} \left(\frac{n}{3} + \frac{1}{3}\right) n \left(n + \frac{1}{2}\right) \\ &= \left(\frac{n}{5} + \frac{1}{5}\right) \left(n + \frac{1}{2}\right) n(n+1)n \\ &\quad - \left(\frac{n}{5} + \frac{1}{5}\right) \left(n + \frac{1}{2}\right) n \cdot \frac{1}{3}. \end{aligned}$$

Ibn al-Haytham stated his result verbally in a form we translate into modern notation as

$$\sum_{i=1}^n i^4 = \left(\frac{n}{5} + \frac{1}{5}\right) n \left(n + \frac{1}{2}\right) \left[(n+1)n - \frac{1}{3} \right].$$

The result can also be written as a polynomial:

$$\sum_{i=1}^n i^4 = \frac{n^5}{5} + \frac{n^4}{2} + \frac{n^3}{3} - \frac{n}{30}.$$

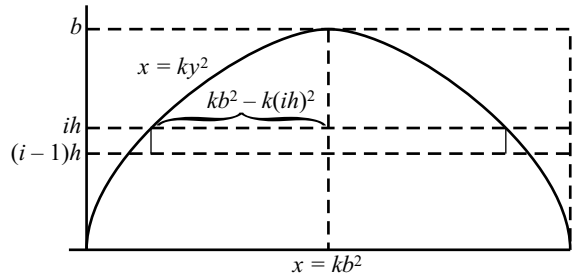


Figure 2.

It is clear that this formula can be used as Fermat and Roberval used Roberval’s inequality to show that

$$\lim_{n \rightarrow \infty} \frac{\sum_{i=1}^n i^4}{n^5} = \frac{1}{5}.$$

Ibn al-Haytham used his result on sums of integral powers to perform what we would call an integration. In particular, he applied his result to determine the volume of the solid formed by rotating the parabola $x = ky^2$ around the line $x = kb^2$ perpendicular to the axis of the parabola, and showed that this volume is $8/15$ of the volume of the cylinder of radius kb^2 and height b . (See Figure 2.) His formal argument was a typical Greek-style exhaustion argument using a double *reductio ad absurdum*, but in essence his method involved slicing the cylinder and paraboloid into n disks, each of thickness $h = b/n$, and then adding up the disks. The i th disk in the paraboloid has radius $kb^2 - k(ih)^2$ and therefore has volume

$$\pi h (kh^2 n^2 - ki^2 h^2)^2 = \pi k^2 h^5 (n^2 - i^2)^2.$$

The total volume of the paraboloid is therefore approximated by

$$\pi k^2 h^5 \sum_{i=1}^{n-1} (n^2 - i^2)^2 = \pi k^2 h^5 \sum_{i=1}^{n-1} (n^4 - 2n^2 i^2 + i^4).$$

But since ibn al-Haytham knew the formulas for the sums of integral squares and fourth powers, he could calculate that

$$\begin{aligned} \sum_{i=1}^{n-1} (n^4 - 2n^2 i^2 + i^4) &= \frac{8}{15} (n-1)n^4 + \frac{1}{30} n \\ &= \frac{8}{15} n \cdot n^4 - \frac{1}{2} n^4 - \frac{1}{30} n \end{aligned}$$

and therefore that

$$\frac{8}{15} (n-1)n^4 < \sum_{i=1}^{n-1} (n^2 - i^2)^2 < \frac{8}{15} n \cdot n^4.$$

But the volume of a typical slice of the circumscribing cylinder is $\pi h((kb)^2)^2 = \pi k^2 h^5 n^4$, and therefore the total volume of the cylinder is

$$\pi k^2 h^5 (n-1)n^5,$$

while the volume of the cylinder less its “top slice” is $\pi k^2 h^5 (n-1)n^4$. The inequality then shows that the volume of the paraboloid is bounded between $8/15$ of the cylinder less its top slice and $8/15$ of the entire cylinder. Because the top slice can be made as small as desired by taking n sufficiently large, it follows that the volume of the paraboloid is exactly $8/15$ of the volume of the cylinder as asserted.

Ibn al-Haytham’s formula for the sum of fourth powers shows up in other places in the Islamic world over the next few centuries. It appears in the work of Abu-l-Hasan ibn Haydur (d. 1413), who lived in what is now Morocco, and in the work of Abu Abdallah ibn Ghazi (1437–1514), who also lived in Morocco. (See [3] for details.) Furthermore, one also finds the formula in *The Calculator’s Key* of Ghiyath al-Din Jamshid al-Kashi (d. 1429), a mathematician and astronomer whose most productive years were spent in Samarkand, now in Uzbekistan, in the court of Ulugh Beg. We do not know, however, how these mathematicians learned of the formula or for what purpose they used it.

3 Trigonometric series in sixteenth-century India

The sum formulas for integral powers surface in sixteenth-century India and they are used to develop the power series for the sine, cosine, and arctangent. These power series appear in Sanskrit verse in the *Tantrasangraha-vyakhya* (of about 1530), a commentary on a work by Kerala Gargya Nilakantha (1445–1545) of some 30 years earlier. Unlike the situation for many results of Indian mathematics, however, detailed derivation of these power series exists, in the *Yuktibhasa*, a work in Malayalam, the language of Kerala, the southwestern region of India. This latter work was written by Jyesthadeva (1500–1610), who credits these series to Madhava, an Indian mathematician of the fourteenth century.

Even though we do not know for sure whether Madhava was the first discoverer of the series, it is clear that the series were known in India long before the time of Newton. But why were the Indians interested in these matters? India had a long tradition of astronomical research, dating back to at least the middle of the first millennium B.C. The In-

dians had also absorbed Greek astronomical work and its associated mathematics during and after the conquest of northern India by Alexander the Great in 327 B.C. Hence the Indians became familiar with Greek trigonometry, based on the chord function, and then gradually improved it by introducing our sine, cosine, and tangent. Islamic mathematicians learned trigonometry from India, introduced their own improvements, and, after the conquest of northern India by a Moslem army in the twelfth century, brought the improved version back to India. (See [4] for more details.)

The interaction of astronomy with trigonometry brings an increasing demand for accuracy. Thus Indian astronomers wanted an accurate value for π (which comes from knowing the arctangent power series) and also accurate values for the sine and cosine (which comes from their power series) so they could use these values in determining planetary positions. Because a recent article [8] in *Mathematics Magazine* discussed the arctangent power series, we will here consider only the sine and cosine series.

The statement of the Indian rule for determining these series is as follows: “Obtain the results of repeatedly multiplying the arc $[s]$ by itself and then dividing by $2, 3, 4, \dots$ multiplied by the radius $[\rho]$. Write down, below the radius (in a column) the even results [i.e., results corresponding to $n = 2, 4, 6, \dots$ in $s^n/n!\rho^{n-1}$], and below the radius (in another column) the odd results [corresponding to $n = 3, 5, 7, \dots$ in $s^n/n!\rho^{n-1}$]. After writing down a number of terms in each column, subtract the last term of either column from the one next above it, the remainder from the term next above, and so on, until the last subtraction is made from the radius in the first column and from the arc in the second. The two final remainders are respectively the cosine and the sine, to a certain degree of approximation.” [6, p. 3] These words can easily be translated into the formulas:

$$\begin{aligned} x = \cos s &= \rho - \frac{s^2}{2!\rho} + \frac{s^4}{4!\rho^3} - \dots \\ &+ (-1)^n \frac{s^{2n}}{(2n)!\rho^{2n-1}} + \dots \\ y = \sin s &= s - \frac{s^3}{3!\rho^2} + \frac{s^5}{5!\rho^4} - \dots \\ &+ (-1)^n \frac{s^{2n+1}}{(2n+1)!\rho^{2n}} + \dots \end{aligned}$$

(These formulas reduce to the standard power series when ρ is taken to be 1.)

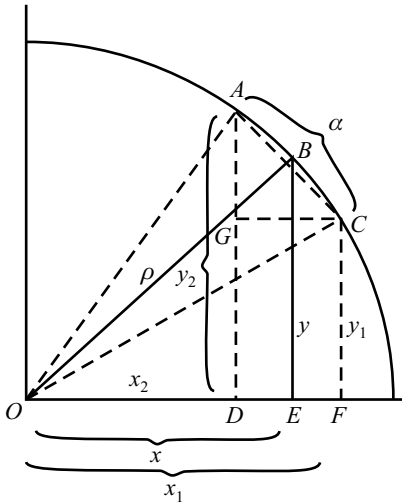


Figure 3.

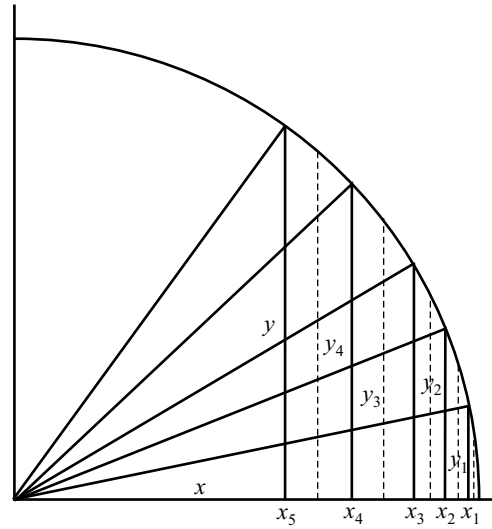


Figure 4.

The Indian derivations of these results begin with the obvious approximations to the cosine and sine for small arcs and then use a “pull yourself up by your own bootstraps” approach to improve the approximation step by step. The derivations all make use of the notion of differences, a notion used in other aspects of Indian mathematics as well. In our discussion of the Indian method, we will use modern notation to enable the reader to follow these sixteenth-century Indian ideas.

We first consider the circle of Figure 3 with a small arc $\alpha = \widehat{AC} \approx AC$. From the similarity of triangles AGC and OEB , we get

$$\frac{x_1 - x_2}{\alpha} = \frac{y}{\rho} \quad \text{and} \quad \frac{y_2 - y_1}{\alpha} = \frac{x}{\rho}$$

or $\frac{\alpha}{\rho} = \frac{x_1 - x_2}{y} = \frac{y_2 - y_1}{x}$.

In modern terms, if $\angle BOF = \theta$ and $\angle BOC = \angle AOB = d\theta$, these equations amount to

$$\begin{aligned} \sin(\theta + d\theta) - \sin(\theta - d\theta) &= \frac{y_2 - y_1}{\rho} = \frac{\alpha x}{\rho^2} \\ &= \frac{2\rho d\theta}{\rho} \cos \theta \\ &= 2 \cos \theta d\theta \end{aligned}$$

and

$$\begin{aligned} \cos(\theta + d\theta) - \cos(\theta - d\theta) &= \frac{x_2 - x_1}{\rho} = -\frac{\alpha y}{\rho^2} \\ &= -\frac{2\rho d\theta}{\rho} \sin \theta \\ &= -2 \sin \theta d\theta. \end{aligned}$$

Now, suppose we have a small arc s divided into n equal subarcs, with $\alpha = s/n$. For simplicity we take $\rho = 1$, although the Indian mathematicians did not. By applying the previous results repeatedly, we get the following sets of differences for the y 's (Figure 4) (where $y_n = y = \sin s$):

$$\begin{aligned} \Delta_n y &= y_n - y_{n-1} = \alpha x_n \\ \Delta_{n-1} y &= y_{n-1} - y_{n-2} = \alpha x_{n-1} \\ &\dots \\ \Delta_2 y &= y_2 - y_1 = \alpha x_2 \\ \Delta_1 y &= y_1 - y_0 = \alpha x_1. \end{aligned}$$

Similarly, the differences for the x 's can be written

$$\begin{aligned} \Delta_{n-1} x &= x_n - x_{n-1} = -\alpha y_{n-1} \\ &\dots \\ \Delta_2 x &= x_3 - x_2 = -\alpha y_2 \\ \Delta_1 x &= x_2 - x_1 = -\alpha y_1. \end{aligned}$$

We next consider the second differences on the y 's:

$$\begin{aligned} \Delta_2 y - \Delta_1 y &= y_2 - y_1 - y_1 + y_0 \\ &= \alpha(x_2 - x_1) = -\alpha^2 y_1. \end{aligned}$$

In other words, the second difference of the sines is proportional to the negative of the sine. But since $\Delta_1 y = y_1$, we can write this result as

$$\Delta_2 y = y_1 - \alpha^2 y_1.$$

Similarly, since

$$\begin{aligned} \Delta_3 y - \Delta_2 y &= y_3 - y_2 - y_2 + y_1 \\ &= \alpha(x_3 - x_2) = -\alpha^2 y_2, \end{aligned}$$

it follows that

$$\Delta_3 y = \Delta_2 y - \alpha^2 y_2 = y_1 - \alpha^2 y_1 - \alpha^2 y_2,$$

and, in general, that

$$\Delta_n y = y_1 - \alpha^2 y_1 - \alpha^2 y_2 - \cdots - \alpha^2 y_{n-1}.$$

But the sine equals the sum of its differences:

$$\begin{aligned} y &= y_n = \Delta_1 y + \Delta_2 y + \cdots + \Delta_n y \\ &= n y_1 - [y_1 + (y_1 + y_2) + (y_1 + y_2 + y_3) \\ &\quad + \cdots + (y_1 + y_2 + \cdots + y_{n-1})] \alpha^2. \end{aligned}$$

Also, $s/n \approx y_1 \approx \alpha$, or $n y_1 \approx s$. Naturally, the larger the value of n , the better each of these approximations is. Therefore,

$$y \approx s - \lim_{n \rightarrow \infty} \left(\frac{s}{n} \right)^2 [y_1 + (y_1 + y_2) + \cdots + (y_1 + y_2 + \cdots + y_{n-1})].$$

Next we add the differences of the x 's. We get

$$x_n - x_1 = -\alpha(y_1 + y_2 + \cdots + y_{n-1}).$$

But $x_n \approx x = \cos s$ and $x_1 \approx 1$. It then follows that

$$x \approx 1 - \lim_{n \rightarrow \infty} \left(\frac{s}{n} \right) (y_1 + y_2 + \cdots + y_{n-1}).$$

To continue the calculation, the Indian mathematicians needed to approximate each y_i and use these approximations to get approximations for $x = \cos s$ and $y = \sin s$. Each new approximation in turn is placed back in the expressions for x and y and leads to a better approximation. Note first that if y is small, y_i can be approximated by is/n . It follows that

$$\begin{aligned} x &\approx 1 - \lim_{n \rightarrow \infty} \left(\frac{s}{n} \right) \left[\frac{s}{n} + \frac{2s}{n} + \cdots + \frac{(n-1)s}{n} \right] \\ &= 1 - \lim_{n \rightarrow \infty} \left(\frac{s}{n} \right)^2 [1 + 2 + \cdots + (n-1)] \\ &= 1 - \lim_{n \rightarrow \infty} \frac{s^2}{n^2} \left[\frac{n(n-1)}{2} \right] \\ &= 1 - \frac{s^2}{2}. \end{aligned}$$

Similarly,

$$\begin{aligned} y &\approx s - \lim_{n \rightarrow \infty} \left(\frac{s}{n} \right)^2 \left[\frac{s}{n} + \left(\frac{s}{n} + \frac{2s}{n} \right) + \cdots \right. \\ &\quad \left. + \left(\frac{s}{n} + \frac{2s}{n} + \cdots + \frac{(n-1)s}{n} \right) \right] \\ &= s - \lim_{n \rightarrow \infty} \frac{s^3}{n^3} [1 + (1+2) + (1+2+3) \\ &\quad + \cdots + (1+2+\cdots+(n-1))] \\ &= s - \lim_{n \rightarrow \infty} \frac{s^3}{n^3} [n(1+2+\cdots+(n-1)) \\ &\quad - (1^2+2^2+\cdots+(n-1)^2)] \\ &= s - s^3 \lim_{n \rightarrow \infty} \left[\frac{\sum_{i=1}^{n-1} i}{n^2} - \frac{\sum_{i=1}^{n-1} i^2}{n^3} \right] \\ &= s - s^3 \left(\frac{1}{2} - \frac{1}{3} \right) \\ &= s - \frac{s^3}{6}, \end{aligned}$$

and there is a new approximation for y and therefore for each y_i . Note that in the transition from the second to the third lines of this calculation the Indians used ibn al-Haytham's equation (*) for the case $k = 1$. Although the Indian mathematicians did not refer to ibn al-Haytham or any other predecessor, they did explicitly sketch a proof of this result in the general case and used it to show that, for any k , the sum of the k th powers of the first n integers is approximately equal to $n^{k+1}/(k+1)$. This result was used in the penultimate line of the above calculation in the cases $k = 1$ and $k = 2$ and in the derivation of the power series for the arctangent as discussed in [8].

To improve the approximation for sine and cosine, we now assume that $y_i \approx (is/n) - (is)^3/(6n^3)$ in the expression for $x = \cos s$ and use the sum formula in the case $k = 3$ to get

$$\begin{aligned} x &\approx 1 - \lim_{n \rightarrow \infty} \frac{s}{n} \left[\frac{s}{n} - \frac{s^3}{6n^3} + \frac{2s}{n} - \frac{(2s)^3}{6n^3} + \cdots \right. \\ &\quad \left. + \frac{(n-1)s}{n} - \frac{((n-1)s)^3}{6n^3} \right] \\ &= 1 - \frac{s^2}{2} + \lim_{n \rightarrow \infty} \frac{s^4}{6n^4} [1^3 + 2^3 + \cdots + (n-1)^3] \\ &= 1 - \frac{s^2}{2} + \frac{s^4}{6} \lim_{n \rightarrow \infty} \frac{\sum_{i=1}^{n-1} i^3}{n^4} \\ &= 1 - \frac{s^2}{2} + \frac{s^4}{6} \cdot \frac{1}{4} \\ &= 1 - \frac{s^2}{2} + \frac{s^4}{24}. \end{aligned}$$

Similarly, ibn al-Haytham’s formula for the case $j = 3$ and the sum formula for the cases $j = 3$ and $j = 4$ lead to a new approximation for $y = \sin s$:

$$\begin{aligned} y &\approx s - \frac{s^3}{6} + \lim_{n \rightarrow \infty} \left(\frac{s}{n}\right)^2 \left[\frac{s^3}{6n^3} + \left(\frac{s^3}{6n^3} + \frac{(2s)^3}{6n^3} \right) \right. \\ &\quad \left. + \cdots + \left(\frac{s^3}{6n^3} + \frac{(2s)^3}{6n^3} + \cdots + \frac{((n-1)s)^3}{6n^3} \right) \right] \\ &= s - \frac{s^3}{6} + \lim_{n \rightarrow \infty} \frac{s^5}{6n^5} [1^3 + (1^3 + 2^3) \\ &\quad + \cdots + (1^3 + 2^3 + \cdots + (n-1)^3)] \\ &= s - \frac{s^3}{6} \\ &\quad + \lim_{n \rightarrow \infty} \frac{s^5}{6n^5} [n(1^3 + 2^3 + \cdots + (n-1)^3) \\ &\quad - (1^4 + 2^4 + \cdots + (n-1)^4)] \\ &= s - \frac{s^3}{6} + \frac{s^5}{6} \lim_{n \rightarrow \infty} \left[\frac{\sum_{i=1}^{n-1} i^3}{n^4} - \frac{\sum_{i=1}^{n-1} i^4}{n^5} \right] \\ &= s - \frac{s^3}{6} + \frac{s^5}{6} \left(\frac{1}{4} - \frac{1}{5} \right) = s - \frac{s^3}{6} + \frac{s^5}{120}. \end{aligned}$$

Because Jyesthadeva considers each new term in these polynomials as a correction to the previous value, he understood that the more terms taken, the more closely the polynomials approach the true values for the sine and cosine. The polynomial approximations can thus be continued as far as necessary to achieve any desired approximation. The Indian authors had therefore discovered the sine and cosine power series!

4 Conclusion

How close did Islamic and Indian scholars come to inventing the calculus? Islamic scholars nearly developed a general formula for finding integrals of polynomials by A.D. 1000 — and evidently could find such a formula for any polynomial in which they were interested. But, it appears, they were not interested in any polynomial of degree higher than four, at least in any of the material which has so far come down to us. Indian scholars, on the other hand, were by 1600 able to use ibn al-Haytham’s sum formula for arbitrary integral powers in calculating power series for the functions in which they were interested. By the same time, they also knew how to calculate the differentials of these functions. So some of the basic ideas of calculus were known

in Egypt and India many centuries before Newton. It does not appear, however, that either Islamic or Indian mathematicians saw the necessity of connecting some of the disparate ideas that we include under the name calculus. There were apparently only specific cases in which these ideas were needed.

There is no danger, therefore, that we will have to rewrite the history texts to remove the statement that Newton and Leibniz invented the calculus. They were certainly the ones who were able to combine many differing ideas under the two unifying themes of the derivative and the integral, show the connection between them, and turn the calculus into the great problem-solving tool we have today. But what we do not know is whether the immediate predecessors of Newton and Leibniz, including in particular Fermat and Roberval, learned of some of the ideas of the Islamic or Indian mathematicians through sources of which we are not now aware.

The entire question of the transmission of mathematical knowledge from one culture to another is a matter of current research and debate. In particular, with more medieval Arabic manuscripts being discovered and translated into European languages, the route of some mathematical ideas can be better traced from Iraq and Iran into Egypt, then to Morocco and on into Spain. (See [3] for more details.) Medieval Spain was one of the meeting points between the older Islamic and Jewish cultures and the emerging Latin-Christian culture of Europe. Many Arabic works were translated there into Latin in the twelfth century, sometimes by Jewish scholars who also wrote works in Hebrew. But although there is no record, for example, of ibn al-Haytham’s work on sums of integral powers being translated at that time, certain ideas he used do appear in both Hebrew and Latin works of the thirteenth century. And since the central ideas of his work occur in the Indian material, there seems a good chance that transmission to India did occur. Answers to the questions of transmission will require much more work in manuscript collections in Spain and the Maghreb, work that is currently being done by scholars at the Centre National de Recherche Scientifique in Paris. Perhaps in a decade or two, we will have evidence that some of the central ideas of calculus did reach Europe from Africa or Asia.

References

1. J. L. Berggren, *Episodes in the Mathematics of Medieval Islam*, Springer-Verlag, New York, 1986.

2. Walter E. Clark, *The Aryabhatiya of Aryabhata*, University of Chicago Press, Chicago, 1930.
3. Ahmed Djebbar, *Enseignement et Recherche Mathématiques dans le Maghreb des XIII^e–XIV^e Siècles* (Publications Mathématiques D'Orsay No. 81-02) Université de Paris - Sud, Orsay, France, 1981.
4. Victor Katz, *A History of Mathematics: An Introduction*, Harper Collins Publishers, New York, 1993.
5. Michael Mahoney, *The Mathematical Career of Pierre de Fermat 1601–65*, Princeton University Press, Princeton, NJ, 1973.
6. C. T. Rajagopal and A. Venkataraman, The sine and cosine power-series in Hindu mathematics, *J. of the Royal Asiatic Society of Bengal - Science* 15 (1949), 1–13.
7. Roshdi Rashed, Ibn al-Haytham et la mesure du paraboloïdes, *J. for the History of Arabic Science* 5 (1981), 262–291.
8. Ranjan Roy, The discovery of the series formula for π by Leibniz, Gregory and Nilakantha, *Mathematics Magazine* 63 (1990), 291–306.
9. D. T. Whiteside, *The Mathematical Papers of Isaac Newton*, Cambridge University Press, Cambridge, 1967–1981.

Was Calculus Invented in India?

DAVID BRESSOUD

College Mathematics Journal 33 (2002), 2–13

1 Introduction

No. Calculus was not invented in India. But two hundred years before Newton or Leibniz, Indian astronomers came very close to creating what we would call calculus. Sometime before 1500, they had advanced to the point where they could apply ideas from both integral and differential calculus to derive the infinite series expansions of the sine, cosine, and arctangent functions:

$$\begin{aligned}\sin x &= x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} - \cdots, \\ \cos x &= 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} - \cdots, \\ \arctan x &= x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} - \cdots.\end{aligned}$$

Roy [13] and Katz [7, 8] have given excellent expositions of the Indian derivation of these infinite summations. I will give a slightly different explanation of how Indian astronomers obtained the sine and cosine expansions, with an emphasis on the succession of problems and insights that ultimately led to these series.

This story provides illuminations of calculus that may have pedagogical implications. The traditional introduction of calculus is as a collection of algebraic techniques that solve essentially geometric problems: calculation of areas and construction of tangents. This was not the case in India. There, ideas of calculus were discovered as solutions to essentially algebraic problems: evaluating sums and interpolating tables of sines.

Geometry was well developed in pre-1500 India. As we will see, it played a role. But it was, at best, a bit player. The story of calculus in India shows us how calculus can emerge in the absence of the

traditional geometric context. This story should also serve as a cautionary tale, for what did emerge was sterile. These mathematical discoveries led nowhere. Ultimately, they were forgotten, saved from oblivion only by modern scholars.

2 Greek origins of trigonometry

Trigonometry arose from, and for over fifteen hundred years was used exclusively for, the study of astronomy/astrology. Hipparchus of Nicæa (*ca.* 161–126 BC) is considered the greatest astronomer of antiquity and the originator of trigonometry. Trigonometry was born in response to a scientific crisis. The Greek attempt to cast astronomy in the language of geometry was running up against the disturbing fact that the heavens are lop-sided. New tools were needed for analyzing astronomical phenomena.

Let me paint the background to this crisis. It begins with the assumption that the earth is stationary. While this was debated in early Greek science—does the earth go around the sun or the sun around the earth?—the simple fact that we perceive no sense of motion is a powerful indication that the earth does not move. In fact, when in the early seventeenth century it became clear that the earth revolves about the sun, it created a tremendous problem for scientists: How to explain how this was possible? How could it be that we were spinning at thousands of miles per hour and hurtling through space at even greater speeds without experiencing any of this? Surely if the earth did move, we would have been flung off long ago. Newton's great accomplishment in the *Principia* was to solve this problem. He created inertial mechanics for this purpose, building it with the then nascent tools of calculus.

So we begin with a fixed and immovable earth.

Above it is the great dome of the night sky, rotating once in every 24 hours. In far antiquity it was realized that the stars do not actually disappear during the day. They are present, but impossible to see against the glare of the sun. The position of the sun in this dome is not fixed. During the year, it travels in its own circle, called the *ecliptic*, through the constellations. One can tell the season by locating the position of the sun in its annual journey around this great circle. This is what the zodiac does. The sign of the zodiac describes the location of the sun by pinpointing the constellation in which it is located.

Most stars are fixed in the rotating dome of the sky, but a few, called the *wanderers* or, in Greek, the *planetes* (hence our word planets), also move across the dome following this same ecliptic circle. If the position of the sun is so important in determining seasons of heat and cold, rain and drought, it appears self-evident that the positions of the wanderers should have important—if more subtle—influences on our lives. Astronomy/astrology was born.

Aristotle, in the 4th century BC, inherited a world-view that saw the earth as the fixed center of the universe with the moon, sun, and planets embedded in concentric, ethereal spheres that rotated with perfect regularity around us. It became the basis for a comprehensive world-view that was tight and consistent and would last for almost two millenia. But its first cracks appeared in less than two hundred years.

The four cardinal points of the great circle traversed by the sun mark the boundaries of the seasons: winter solstice, spring equinox, summer solstice, and autumnal equinox. If the sun travels the

ecliptic at constant speed, the four seasons should be of equal length. They are not (see Figure 1). Winter solstice to spring equinox is a short 89 days. Spring is almost 90 days. Summer, the longest season, is over $93\frac{1}{2}$ days. And fall comes close to 93 days. If, in fact, the sun moves at a constant speed, this can only mean that the earth is off-center. Hipparchus tackled the problem of calculating the position of the earth.

The basic problem of trigonometry as understood by Hipparchus and his contemporaries is the following: Given an arc of a circle, find the length of the chord that connects the endpoints of that arc (see Figure 2). This chord length depends on both the length of the arc and the radius of the circle. For the Greeks, as for all scientists right through Newton, 90° was not the measure of a right angle, but of the distance around one quarter of the circumference of a circle. Degrees were a measure of distance. Given a circle of circumference 360° , it would be natural to take the radius to be $360/2\pi = 57.2957795\dots$ For greater accuracy, the circumference of this standard circle could be measured in minutes. The circumference is then 21,600 minutes and the radius is 3437.74677\dots It would become common in Indian trigonometry to use a radius of 3438. There is some evidence that Hipparchus, whose trigonometric tables no longer survive, also may have used a radius of 3438.

Hipparchus was probably the first to construct a table of values of the length of the chord for a given arc, what is sometimes called $\text{crd } \alpha$. In modern trigonometric notation, the chord is twice the sine of

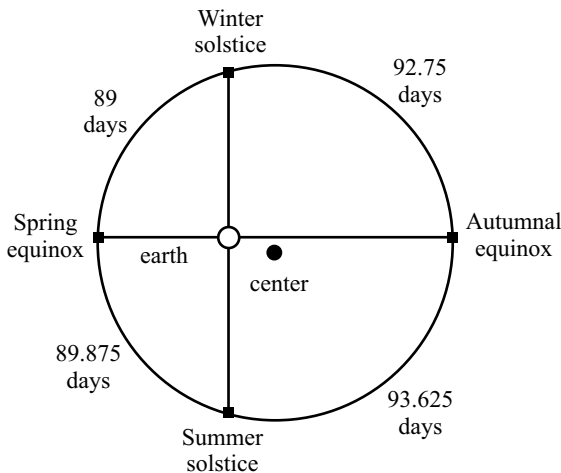


Figure 1. The unequal seasons, rounded to nearest 1/8 day.

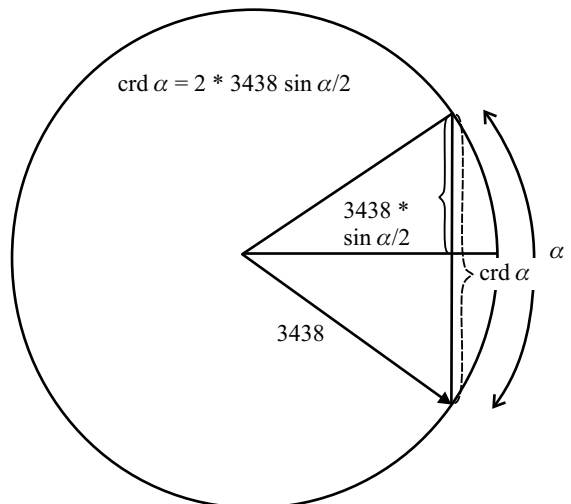


Figure 2. The relationship between $\text{crd } \alpha$ and $\sin \alpha$.

half the angle, multiplied by the radius of the circle which we will take to be 3438 (see Figure 2):

$$\text{crd } \alpha = 3438 \cdot 2 \sin(\alpha/2).$$

For the problem of the position of the earth, the arc from the winter solstice to the summer solstice is approximately $176^\circ 18'$. Assuming that we know that $\text{crd } 176^\circ 18' = 6872'$, it follows that half of the chord is $3436'$. We can now use the Pythagorean theorem to find the distance from the center of the circle to this chord:

$$\text{distance} = \sqrt{3438^2 - 3436^2} \approx 117'.$$

This chord is 117 minutes, almost two full degrees, off center.

Over succeeding centuries, as astronomical observations became more accurate, the model for the movement of sun and planets became more complicated. Planets will seem to pause and reverse direction.¹ This was explained by putting small spheres inside each crystal ring, epicycles on which each planet would rotate around a point which itself traveled around the earth. Even with an off-centered earth, it was necessary to vary the speed of the spheres. This was often accomplished by adding an equant, a point from which the angular velocity of the center of the small circle appears constant (see Figure 3). All of the workings of this model relied on trigonometric calculations, and these calculations relied on an accurate table of chords.

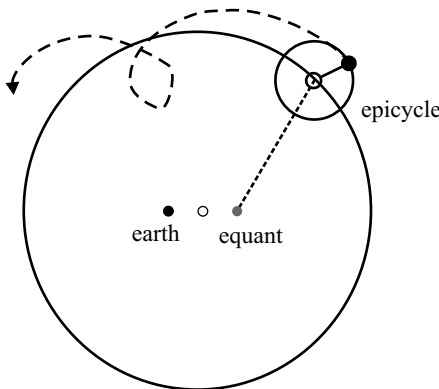


Figure 3. An epicycle combined with an equant. The planet circles the center of the epicycle. The center of the epicycle moves so that its angular velocity relative to the equant is constant.

¹See <http://alpha.lasalle.edu/~smithsc/Astronomy/retrograd.html> for an illustration and explanation of retrograde motion.

By the end of the first century AD, Menelaus of Alexandria knew the formulas for the chords of sums and differences of angles, double and half angles. With these, he was able to construct an accurate table of chords. In the second century, Ptolemy, also of Alexandria, published his system of the heavens, including a table of chords for angles in increments of half a degree, equivalent to a table of sines in increments of a quarter-degree. It is important to our story to look at how this table was constructed. While it was given as a table of chords, I will explain it in terms of more familiar sines.

Beginning with the fact that $\sin 30^\circ = \frac{1}{2}$ and using the half-angle formula

$$\sin \alpha = \sqrt{\frac{1 - \cos(2\alpha)}{2}},$$

one can calculate the sines of 15° , $7^\circ 30'$, and $3^\circ 45'$. Going back at least to Archimedes, it was known that

$$\sin 36^\circ = \frac{1}{2} \sqrt{\frac{5 - \sqrt{5}}{2}},$$

and so we get the sine of 18° . Using the sines of 15° and 18° and the difference of angles formula,

$$\sin(\alpha - \beta) = \sin \alpha \cos \beta - \cos \alpha \sin \beta,$$

we get the sine of 3° . Now we can calculate the sines of $1^\circ 30'$ and $45'$.

We are down to a very small angle. The Greeks knew that for very small angles, we have the approximation

$$\frac{\sin \alpha}{\sin \beta} \approx \frac{\alpha}{\beta}.$$

It follows that

$$\sin 1^\circ \approx \frac{4}{3} \sin \frac{3^\circ}{4}.$$

This is not a bad approximation. The error is of the same order of magnitude as that introduced by using 3438 as the radius of a circle of circumference 21600, less than 1 part in 10000. Two more iterations of the half angle formula, and we are down to the sine of $15'$. Now we can use the sum and difference of angles formulas to fill in the missing values in the table.

Ptolemy's *Almagest* was the last great scientific achievement of the Græco-Roman world. Fortunately, India was just coming into its high classical period. Indian astronomers learned of the Greek accomplishments and began to incorporate them into

their own science.² They did not stop with borrowing Greek ideas. They began to improve on what the Greeks had accomplished. Among their improvements would be conceptual breakthroughs that would allow them to reduce the errors to 1 part in 10¹².

3 Trigonometry in classical India

One of the first innovations was to work with the half-chord rather than the Greek chord, what was called the *ardha-jyā* or “half bowstring,” eventually simplified to just *jyā* (bowstring) or *jīvā*. Islamic astronomers learned much of their trigonometry from India; Europe would learn it from North Africa. That is why today we use sines instead of chords.³ But the greatest contribution to trigonometry to come out of India was the analysis of how to interpolate the tables of sines. From this would come the power series for the sine and cosine.

Āryabhata, born in 476, analyzed a fourth century Sanskrit table of sines and described an interesting pattern when he took differences of consecutive entries, and then differences of those differences:

α	3438 $\sin \alpha$	1st difference	2nd difference
3°45'	225		
7°30'	449	224	-2
11°15'	671	222	-3
15°	890	219	-4
18°45'	1105	215	-5
22°30'	1315	210	-5
26°15'	1520	205	-6
30°	1719	199	

Āryabhata observed that these second differences are very close to the value in the second column divided by 225:

$$[\sin(x + 225') - \sin x] - [\sin x - \sin(x - 225')] \approx \frac{-\sin x}{225}.$$

²According to Neugebauer and Pingree [9], the Paulisa-siddhanta and the Romaka-siddhanta (4th century or earlier) are based on Greek astronomical works. Similarities in terminology, calculations, and choices of constants—as well as the names of these works—argue for the importation of Greek astronomical techniques.

³According to Datta and Singh [3], Arab mathematicians used the term *jiba*, clearly derived from the Sanskrit *jīvā*. When Gherardo of Cremona (ca. 1150) translated this into Latin, he misread it as *jaib* which is Arabic for “bosom” or “bay,” and translated it as *sinus* from which we get *sine*.

Datta and Singh [3, pp. 75–77] argue that this could have been derived from the trigonometric identity

$$[\sin(x + \alpha) - \sin x] - [\sin x - \sin(x - \alpha)] = -\sin x \left(\frac{2 \sin(\alpha/2)}{3438} \right)^2, \quad (1)$$

in which the argument of the sine is measured in minutes. This derivation is pure speculation, but it does illustrate how Āryabhata’s successors might have come to discover the second derivative formula for the sine. The sum of angles and half-angle formulas that are needed to derive (1) were certainly known by 1200 and probably long before that. Ever since the inception of trigonometry, it had been known that $2 \sin(\alpha/2)/\alpha$ is approximately 1 for small values of α . When the argument of the sine function is measured in minutes, it follows from (1) that

$$\lim_{\alpha \rightarrow 0} \frac{\sin(x + \alpha) - 2 \sin x + \sin(x - \alpha)}{\alpha^2} = \frac{-\sin x}{3438^2} \lim_{\alpha \rightarrow 0} \left(\frac{2 \sin(\alpha/2)}{\alpha} \right)^2 = \frac{-\sin x}{3438^2}.$$

By 665, Brahmagupta of Bhillamāla (modern Bhinmal) in Rajasthan had found the formula that showed how to use the second differences to approximate interpolated values. We assume that we want to find the value of $\sin(x + \epsilon)$ where x is the nearest angle for which we know $\sin x$. We also assume that α is the common difference between angles in our table, so that we also know the sines of $x + \alpha$ and $x - \alpha$. These can be used to approximate the first and second derivatives of $\sin x$:

$$\frac{d}{dx} \sin x \approx \frac{\sin(x + \alpha) - \sin(x - \alpha)}{2\alpha},$$

$$\frac{d^2}{dx^2} \sin x \approx \frac{\sin(x + \alpha) - 2 \sin x + \sin(x - \alpha)}{\alpha^2}.$$

Brahmagupta stated that

$$\begin{aligned} \sin(x + \epsilon) &\approx \sin x + \epsilon \frac{\sin(x + \alpha) - \sin(x - \alpha)}{2\alpha} \\ &\quad + \frac{\epsilon^2}{2} \frac{\sin(x + \alpha) - 2 \sin x + \sin(x - \alpha)}{\alpha^2}. \end{aligned}$$

It is worth noting that this formula is valid no matter what units—degrees, minutes, or radians—we use to measure ϵ and α . We do, however, have to use the same units for both.

What Brahmagupta had discovered is the quadratic case of the Newton interpolation formula.⁴ The right side is the unique quadratic polynomial that agrees with the sine at $x - \alpha$, x , and $x + \alpha$. Note that if α and ϵ are measured in radians and we take the limit as $\alpha \rightarrow 0$, then we get the familiar Taylor polynomial in ϵ :

$$\sin(x + \epsilon) \approx \sin x + \epsilon \cos x - \frac{\epsilon^2}{2} \sin x.$$

In the early ninth century, Govindasvāmin of Kerala showed how to extend Brahmagupta’s quadratic formula to interpolation formulas for higher powers.

By the twelfth century, Bhāskara II was using the fact that the first difference of the sine, $\sin(x + \epsilon) - \sin x$, is close to $\epsilon \cos x$ in the sense that their ratio approaches 1 as ϵ approaches 0. He also used $(-\sin x)\epsilon^2$ to approximate the second difference of the sine. Around 1400 in a commentary on the work of Govindasvāmin, Parameśvara used the limits of the first, second, and third differences to give a cubic approximation for $\sin(x + \epsilon)$ when $\sin x$ is known:

$$\sin(x + \epsilon) = \sin x + \frac{\epsilon}{R} \cos x - \frac{\epsilon^2}{2R^2} \sin x - \frac{\epsilon^3}{4R^3} \cos x,$$

where the arguments of the trigonometric functions are measured in units equal to R^{-1} radians. This formula is not quite correct. The last denominator should be $6R^3$. But the Indian astronomers were on their way to the general Maclaurin expansions of the sine and cosine.

The exact date and attribution of the series for sine, cosine, and arctangent are uncertain. The earliest unquestioned appearance of these series is in the *Yuktibhāṣā* written by Jyeṣṭhadeva in the early 1500s. Jyeṣṭhadeva based his work on the *Tantrasamgraha* of Nīlakanṭha, written in 1501. A commentary on the *Tantrasamgraha* by one of Jyeṣṭhadeva’s students, the *Tantrasamgraha-vyākhyā*, written prior to 1550, has led Rajagopal and Rangachari [11] to argue that Nīlakanṭha was familiar with these series and that they were part of the oral tradition that accompanied his work. The series for the sine does appear in one of Nīlakanṭha’s later works, the *Āryabhatīya-bhāṣya*, written prior to 1545, where he attributes it to Mādhava who lived approximately 1349–1425. There is additional evidence from the results that Mādhava is known to have authored that he probably did know these se-

ries.⁵ What this all means is that the date of discovery of these series cannot be pinned down any more accurately than after 1350 and before 1550, with evidence suggesting the earlier rather than the later part of this window. In any event, they were discovered in India well over a century before their rediscovery in Europe.

4 The power series expansion for sine

Up to this point, I have translated the Indian formulas into more familiar sines and cosines, but to do proper justice to the Indian derivation of the sine series, I need to state and follow the proof in something closer to the original notation. I will use $jy\bar{a} \alpha$ and $koj \alpha$ (for $kotijy\bar{a}$) to denote, respectively, the half-chord of the arclength α and the half-chord of the complementary angle. Note that these are also dependent on the radius, R . If the sine and cosine are functions of angles measured in radians, then

$$\begin{aligned} jy\bar{a} \alpha &= R \sin(\alpha/R), \\ koj \alpha &= R \cos(\alpha/R). \end{aligned}$$

We will present Jyeṣṭhadeva’s proof that

$$jy\bar{a} \alpha = \alpha - \frac{\alpha^3}{R^2 3!} + \frac{\alpha^5}{R^4 5!} - \frac{\alpha^7}{R^6 7!} + \dots$$

The first step is to find the limiting formulas for the first difference of the $jy\bar{a}$ and $kotijy\bar{a}$. In Figure 4, we let \widehat{PX} be the arclength α and \widehat{PR} be $\Delta\alpha$, the change in α . The problem is to estimate $RS = \Delta(jy\bar{a} \alpha) = jy\bar{a}(\alpha + \Delta\alpha) - jy\bar{a} \alpha$ and $PS = \Delta(koj \alpha)$. We mark Q , the midpoint of arc \widehat{PR} , and note that OQ is the perpendicular bisector of chord PR .

For a small change in arclength, the chord PR is a very good approximation to the arc \widehat{PR} , and so we will not distinguish between them. Also, BQ equals $jy\bar{a}(\alpha + \frac{1}{2}\Delta\alpha)$ which we will identify with $jy\bar{a} \alpha$. Similarly, we treat OB as equal to $koj \alpha$. Triangle RSP is similar to triangle OBQ , and therefore

$$\frac{RS}{PR} = \frac{OB}{OQ} \implies \Delta(jy\bar{a} \alpha) = \frac{(\Delta\alpha) koj \alpha}{R}, \quad (2)$$

$$\frac{PS}{PR} = \frac{BQ}{OQ} \implies \Delta(koj \alpha) = \frac{-(\Delta\alpha) jy\bar{a} \alpha}{R}. \quad (3)$$

⁴Newton’s interpolation formula appears in his *Principia Mathematica*. Brook Taylor used it to derive the Taylor series.

⁵See the analyses by Pingree [10] and Sarma [15].

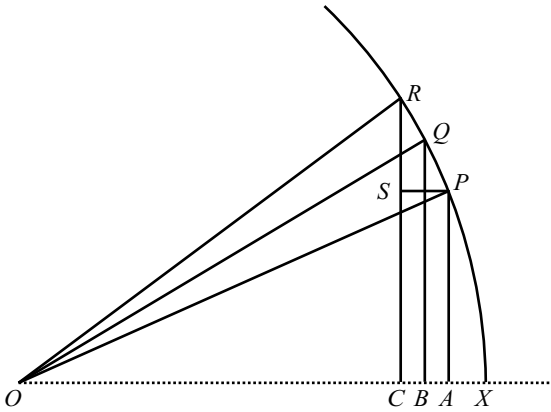


Figure 4. $RS = \Delta(\text{jy}\bar{a} \alpha)$ and $PS = \Delta(\text{koj} \alpha)$.

In modern terms, Jyeṣṭhadeva’s next step is to observe that

$$\sin \alpha = \int_0^\alpha \cos x \, dx,$$

and

$$\cos \alpha = 1 - \int_0^\alpha \sin x \, dx.$$

Using these equalities, a polynomial approximation to the sine can be turned into an approximation of the cosine with degree one higher. A polynomial approximation of the cosine can be turned into an approximation of the sine with degree one higher. We then iterate this process to generate the infinite series.

Let me put the preceding paragraph into notation that is still modern but closer to the spirit of Jyeṣṭhadeva’s construction. He sets $\alpha = n \Delta\alpha$ and observes that the sum of small differences is equal to the large difference:

$$\text{koj} \alpha - \text{koj} 0 = \sum_{i=0}^{n-1} \Delta(\text{koj} (i \Delta\alpha)).$$

He uses (3) and the approximation $\text{jy}\bar{a} \alpha \approx \alpha$ to simplify this:

$$\begin{aligned} \text{koj} \alpha - \text{koj} 0 &= \sum_{i=0}^{n-1} \frac{-(\Delta\alpha)\text{jy}\bar{a} (i \Delta\alpha)}{R} \\ &= \frac{-(\Delta\alpha)^2}{R} \sum_{i=0}^{n-1} i \\ &= \frac{-(\Delta\alpha)^2(n^2 - n)}{2R}. \end{aligned} \tag{4}$$

We know that $\text{koj} 0 = R$, $n\Delta\alpha = \alpha$, and $(\Delta\alpha)^2n$ can be made arbitrarily small by taking $\Delta\alpha$ suffi-

ciently small. Taken with (4), this implies that

$$\text{koj} \alpha \approx R - \frac{\alpha^2}{2R}. \tag{5}$$

We now use this approximation and the result given in (4) to improve the approximation to $\text{jy}\bar{a} \alpha$:

$$\begin{aligned} \text{jy}\bar{a} \alpha - \text{jy}\bar{a} 0 &= \sum_{i=0}^{n-1} \Delta(\text{jy}\bar{a} (i \Delta\alpha)) \\ &= \sum_{i=0}^{n-1} \frac{\Delta\alpha}{R} \left(R - \frac{(i \Delta\alpha)^2}{2R} \right) \\ &= n\Delta\alpha - \frac{(\Delta\alpha)^3}{2R^2} \sum_{i=0}^{n-1} i^2. \end{aligned} \tag{6}$$

We use the fact that $\sum_{i=0}^{n-1} i^2$ is $n^3/3$ plus lower order terms to get the improved approximation:

$$\text{jy}\bar{a} \alpha \approx \alpha - \frac{\alpha^3}{2 \cdot 3 R^2}. \tag{7}$$

In the next iteration we need to know that $\sum_{i=0}^{n-1} i^3$ is $n^4/4$ plus lower order terms. For the general iterative step, we need to know that

$$\sum_{i=0}^{n-1} i^k = \frac{n^{k+1}}{k+1} + \text{lower order terms}. \tag{8}$$

Today we recognize that $\sum_{i=0}^{n-1} (i/n)^k (1/n)$ is a Riemann sum for $\int_0^1 x^k \, dx$. In other words, what we need to know is that

$$\int_0^\alpha x^k \, dx = \frac{\alpha^{k+1}}{k+1}.$$

Jyeṣṭhadeva’s argument for (8) is given by Roy [13]. Katz [8] describes al-Haytham’s derivation of (8) for $k \leq 4$, an approach that is easily extended to any value of k . Al-Haytham lived in eleventh century Egypt, but knowledge of his results may have traveled to India. In fact, this asymptotic estimate for the summation of the k th powers became widely known in the Middle East and India before the fifteenth century. I shall describe the approach used by Nārāyaṇa in his *Gaṇitakaumudī*, written in 1356.

Nārāyaṇa built on earlier observations that

$$\sum_{i=1}^n \binom{i}{1} = \binom{n+1}{2},$$

$$\sum_{i=1}^n \binom{i+1}{2} = \binom{n+2}{3},$$

$$\sum_{i=1}^n \binom{i+2}{3} = \binom{n+3}{4}.$$

The first two of these are ancient and can be found in both Greek and early Jain mathematics. They lend themselves to geometric proofs. Nārāyaṇa's greatest accomplishment was to view these not as geometric, but as formulaic or algebraic (though, of course, he did not have the advantages of our notation). He thought of them as iterated sums. This suggested the following generalization:

$$\sum_{i=1}^n \binom{i+k-1}{k} = \binom{n+k}{k+1}. \quad (9)$$

Each iterated sum, $\binom{n+k-1}{k}$, is equal to

$$\begin{aligned} & \frac{n(n+1)(n+2) \cdots (n+k-1)}{k!} \\ &= \frac{n^k}{k!} + \text{lower order terms.} \end{aligned}$$

It follows from (9) that

$$\begin{aligned} & \sum_{i=1}^n \left(\frac{i^k}{k!} + \text{lower order terms} \right) \\ &= \frac{n^{k+1}}{(k+1)!} + \text{lower order terms,} \end{aligned}$$

which implies (8). It is worth noting that Nārāyaṇa also showed how to use equation (9) to find sums of other specific polynomials in i by first expressing the polynomial as a linear combination of these binomial coefficients.

5 Conclusion

There is no evidence that the Indian work on series was known beyond India, or even outside Kerala, until the nineteenth century. Gold and Pingree [4] assert that by the time these series were rediscovered in Europe, they had, for all practical purposes, been lost to India. The expansions of the sine, cosine, and arc tangent had been passed down through several generations of disciples, but they remained sterile observations for which no one could find much use.

No. Calculus was not discovered in India. I am left wondering how much important mathematics is today known but not yet discovered, passed among a coterie of tightly knit disciples as an intriguing yet seemingly useless insight, lacking the context, the fertilizing connections, that would enable it to blossom and produce its fruit.

Bibliography

1. D. M. Bose, S. N. Sen, and B. V. Subbarayappa, *A Concise History of Science in India*, Indian National Science Academy, 1971.
2. B. Datta and A. N. Singh, revised by K. S. Shukla, Hindu geometry, *Indian Journal of History of Science* 15 (1980) 121–188.
3. —, Hindu trigonometry, *Indian Journal of History of Science* 18 (1983) 39–108.
4. D. Gold and D. Pingree, A hitherto unknown Sanskrit work concerning Madhava's derivation of the power series for sine and cosine, *Historia Scientiarum* 42 (1991) 49–65.
5. R. C. Gupta, An Indian form of third order Taylor series approximation of the sine, *Historia Mathematica* 1 (1974) 287–289.
6. T. Heath, *A History of Greek Mathematics*, reprint of Oxford edition of 1921, Dover, 1981.
7. V. J. Katz, *A History of Mathematics: an Introduction*, 2nd edition, Addison-Wesley, 1998.
8. —, Ideas of calculus in Islam and India, *Math. Magazine* 68 (1995) 163–174.
9. O. Neugebauer and D. Pingree, *The Pañcasiddhāntikā of Varāhamihira*, Det Kongelige Danske Videnskaberne Selskab, Historisk-Filosofiske Skrifter, Vol. 6, Nos. 1 & 2, 1970.
10. D. Pingree, *Jyotiḥśāstra, Astral and Mathematical Literature*, A History of Indian Literature, Vol. 6, Otto Harrassowitz, Weisbaden, 1981.
11. C. T. Rajagopal and M. S. Rangachari, On an untapped source of medieval Keralese mathematics, *Archive for History of Exact Sciences* 18 (1978) 89–102.
12. C. T. Rajagopal and A. Venkataram, The sine and cosine power-series in Hindu mathematics, *J. Royal Asiatic Society of Bengal, Science* 15 (1949) 1–13.
13. R. Roy, The discovery of the series formula for π by Leibniz, Gregory and Nīlakaṇṭha, *Math. Magazine* 63 (1990) 291–306.
14. T. A. Saraswathi, The development of mathematical series in India after Bhaskara II, *Bulletin of the National Institute of Sciences* 21 (1963) 320–343.
15. K. V. Sarma, *A History of the Kerala School of Hindu Astronomy*, Vishveshvaranand Institute, Hoshiarpur, 1972.

An Early Iterative Method for the Determination of $\sin 1^\circ$

FARHAD RIAHI

College Mathematics Journal 26 (1995), 16–21

1 Background

In his popular *History of Mathematics*, Carl B. Boyer [5] dated the medieval period in Europe from 529 A.D. to 1436. It was in 529 that the Byzantine emperor Justinian, fearing a threat to orthodox Christianity, ordered all pagan philosophical schools at Athens to be closed and the scholars dispersed. Rome, then ruled by the Goths, was hardly a hospitable home for the learned, but many found a haven in Sassanide Persia. To Boyer the year 1436 marked the dawn of a new mathematical era in Christian Europe for two reasons. It saw the birth of the most influential European mathematician of the fifteenth century, Johann Mueller, better known as Regiomontanus, and Boyer took 1436 as the probable year of death of al-Kashi, the last in a long lineage of prominent Muslim scholars (who actually died in 1429).

Until recently, historical accounts esteemed Muslim scientists mainly for holding Greek learning in cold storage until Europe was ready to accept it, and indeed the decline of Muslim scholarship did coincide with Europe's emergence from the Middle Ages. But between 750 and 1450, Islamic civilization in fact produced a series of remarkable mathematicians who, among other accomplishments, invented the decimal system (including decimal fractions), created algebra, systematized plane and spherical trigonometry, made important discoveries in these sciences, and developed ingenious methods for solving algebraic equations. Only recently have researchers in the history of mathematics begun to re-discover what medieval and renaissance scholars knew well, the intellectual legacy bequeathed by Muslim scientists. One eminent mathematical historian has lamented this long neglect as

“unfortunate, not only from a scholarly point of view, but from a pedagogical one as well, for Islam's contributions include some gems of mathematical reasoning, accessible to anyone who has learned high school mathematics” [3]. A recent general survey by Victor Katz [10] gives proper weight to the achievements of medieval Muslim mathematicians.

Who was al-Kashi, this Janus-faced mathematician who looked back at the old and anticipated the new? He was born in the second half of the fourteenth century in the town of Kashan (Iran), whence his name Kashani—or al-Kashi, as he is better known in the West. How he learned mathematics is obscure. By his own account, he first led the precarious life of a wandering scholar, seeking patronage at the courts of local lords and dedicating scientific treatises in return. Then, around 1420, he was invited to Samarkand (in present day Uzbekistan) by the Great Khan Ulugh Beg (1393–1449) to help design and construct a state-of-the-art observatory. Al-Kashi remained in Samarkand as the director of this observatory until his death on June 22, 1429.

The author of several important works in mathematics and astronomy, al-Kashi also invented astronomical instruments, such as the planetary equatorium, and perfected existing ones. In the treatise reviewed below, he calculated the value of $\sin 1^\circ$ to a high degree of accuracy by an iterative procedure. In another work, he determined the value of π up to 17 correct decimal digits. His magnum opus, the *Key to Arithmetic* (1427) is a magisterial compendium of arithmetic and algebra which remained a standard textbook in the Muslim world until the seventeenth century, and was probably known in Europe as well.

(Two Byzantine manuscripts on arithmetic exercises brought to Vienna in 1562 refer to his text; see [8].) Al-Kashi's book was the first complete treatment of the theory and application of decimal fractions, which he recommended in place of the sexagesimal system [3], [10]—the standard scientific computational scheme since Ptolemy (second century A.D.)

2 Al-Kashi's determination of $\sin 1^\circ$

Al-Kashi's original treatise about chords and sines is unfortunately lost. In the preface to his *Key to Arithmetic*, he mentions having written "a treatise on chords and sines for use in the calculation of the chord and sine of the third of an arc whose chord or sine is known, . . . one of the problems that my predecessors have found difficult" to solve [9]. Fortunately, his method survived in brief accounts by a colleague at the Samarkand observatory and by this astronomer's grandson. Their versions, translated from Arabic (the former *lingua franca* of the Muslim world) into French, English, and Russian, have prompted several scholarly studies [2]. Yet within a commentary in Persian written by another colleague of al-Kashi named Abd-el Ali Birjandi [4], there exists a more complete description of al-Kashi's procedure, apparently unknown to Western scholars. My article is based on Birjandi's account. Using modern symbols, I shall describe al-Kashi's ingenious iterative method, establish the existence and uniqueness of a solution, and show the convergence of this algorithm, since none of these issues is rigorously addressed in the studies mentioned above.

To appreciate the importance of al-Kashi's contribution, bear in mind that trigonometry assumed a central place in Muslim mathematics. It supplied the tools for accurate astronomical calculations, the elaboration of calendars, geographical measurements, and navigation. Based on translations of the Indian *Surya Siddhanta* (fourth or fifth century A.D.), the *Spherica* of Menelaus (first century A.D.), and especially Ptolemy's *Almagest*, Muslim mathematicians developed plane and spherical trigonometry to an advanced level. They promoted the use of the modern trigonometric functions instead of the chord functions; proved the half-angle formulas, the sine law, and the addition theorem for sines; developed linear and quadratic interpolation procedures; and established tables for the values of the trigonometric

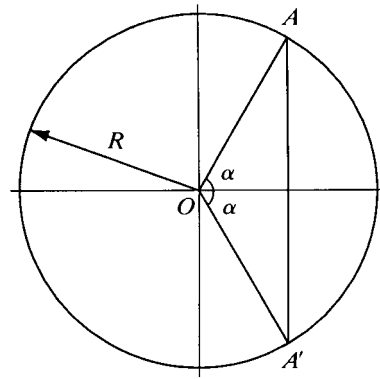


Figure 1. $AA' = \text{crd } 2\alpha = 2R \sin \alpha$

functions.

In this massive labor of computing trigonometric tables, knowing the precise value of $\sin 1^\circ$ is of fundamental importance. From the value of $\sin 1^\circ$ and the values of a few other basic sines, the trigonometric formulas generate $\sin p^\circ$ for all integer values of p . The half-angle formulas can then be used to compute the sines in intervals of $\frac{1}{2}^\circ$ and $\frac{1}{4}^\circ$. Finally, interpolation algorithms yield values for finer subdivisions. Although Ptolemy's interpolation method [1] for the approximate calculation of chords, in particular the chord of 1° , had been refined and used by Muslim mathematicians, they knew well the inherent limitations of this procedure, which quickly grows cumbersome and whose accuracy is restricted by the very inequalities from which it proceeds. To find a simpler and rapidly converging method for the evaluation of $\sin 1^\circ$ was highly desirable, and this became al-Kashi's goal. (See Figure 1 for the relation between $\text{crd } 2\alpha$, the chord of angle 2α , and $\sin \alpha$.)

He begins by setting up an equation expressing $\sin 1^\circ$ in terms of $\sin 3^\circ$. His method is purely geometrical and general, so that he can be justly considered the first to derive the well-known formula

$$\sin 3\phi = 3 \sin \phi - 4 \sin^3 \phi$$

otherwise attributed to Viète in the late sixteenth century.

Consider a semicircle of radius R with center O and diameter AE (see Figure 2). Let B, C, D be points on this semicircle such that $\text{arc } AB = \text{arc } BC = \text{arc } CD$. Ptolemy's theorem [1] applied to the inscribed quadrilateral $ABCD$ yields

$$AB \cdot CD + BC \cdot AD = AC \cdot BD.$$

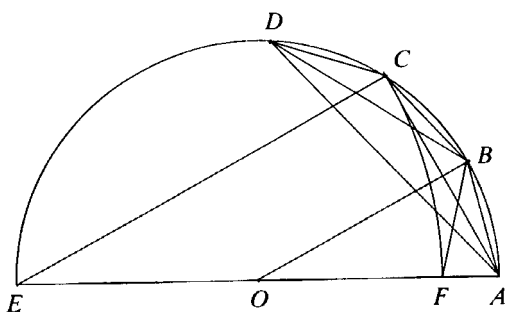


Figure 2.

Since $AB = CD = BC$ and $BD = AC$, we have

$$AB^2 + BC \cdot AD = AC^2. \tag{1}$$

Now determine point F on AE such that $EF = EC$ and consider the similar isosceles triangles ABF and ABO . We have

$$\frac{AB}{AF} = \frac{AO}{AB} \quad \text{or} \quad AF = \frac{AB^2}{R}$$

so that $EF = 2R - AF = 2R - AB^2/R$. On the other hand, in the right triangle AEC ,

$$AC^2 = AE^2 - EC^2 = 4R^2 - EF^2,$$

so that

$$\begin{aligned} AC^2 &= 4R^2 - \left(2R - \frac{AB^2}{R}\right)^2 \\ &= 4AB^2 - \frac{AB^4}{R^2}. \end{aligned} \tag{2}$$

Equations (1) and (2) yield

$$AB^2 + AB \cdot AD = 4AB^2 - \frac{AB^4}{R^2},$$

that is,

$$AD = 3AB - \frac{AB^3}{R^2}. \tag{3}$$

Clearly, if arc $AB = 2\alpha$, then arc $AD = 6\alpha$, and recalling that $\text{crd } 2\alpha = 2R \sin \alpha$ (see Figure 1), equation (3) yields

$$2R \sin 3\alpha = 6R \sin \alpha - \frac{8R^3 \sin^3 \alpha}{R^2}$$

or

$$\sin 3\alpha = 3 \sin \alpha - 4 \sin^3 \alpha.$$

In particular, for $\alpha = 1^\circ$ and setting $\sin 1^\circ = x$, al-Kashi obtains the cubic equation

$$x = \frac{4}{3}x^3 + \frac{1}{3} \sin 3^\circ, \tag{4}$$

one of whose roots is $\sin 1^\circ$. Note that al-Kashi ignored negative roots, since negative numbers had not yet been introduced in his time.

Standard Euclidean constructions, along with accurate algorithms for the extraction of square roots, provide the value of $\sin 30^\circ$ as half the length of the side of a regular hexagon inscribed in the unit circle, and the value of $\sin 36^\circ$ as half the length of the side of a similarly inscribed pentagon. Hence, $\sin 3^\circ = \sin(18^\circ - 15^\circ)$ can be calculated with any desired accuracy.

To calculate the root of equation (4), al-Kashi devised an iterative method known today as *fixed point iteration*. He appears to have used the sexagesimal system, but for clarity I shall render his argument using decimal expansions. Being aware of the inequality $\sin \phi > (1/n) \sin(n \cdot \phi)$, which implies that $\sin 1^\circ > \frac{1}{3} \sin 3^\circ$, he considers the cubic term in equation (4) as a small correction to be added to $\frac{1}{3} \sin 3^\circ$. He then argues as follows:

The root x cannot be substantially larger than $\frac{1}{3} \sin 3^\circ$, so if one considers the decimal expansion of x , the first two or even three digits after the decimal point should be identical with those in the decimal expansion of $\frac{1}{3} \sin 3^\circ$. Al-Kashi uses the known value of $\sin 3^\circ$, so that $\frac{1}{3} \sin 3^\circ = 0.0174453\dots$, and concludes that the root should have a decimal expansion $x = 0.01a_1a_2a_3\dots$ where the a_k are whole numbers between 0 and 9. Inserting this first estimate into equation (4), he finds

$$\begin{aligned} 0.010a_1a_2a_3\dots &= \frac{4}{3}(0.01a_1a_2a_3\dots)^3 \\ &\quad + \frac{1}{3} \sin 3^\circ \end{aligned} \tag{5}$$

or (subtracting 0.01 from both sides)

$$\begin{aligned} 0.00a_1a_2a_3\dots &= \frac{4}{3}(0.01a_1a_2a_3\dots)^3 \\ &\quad + 0.0074453\dots \end{aligned}$$

The last equality should hold true digit by digit, and since the cubic term in the right-hand side has its first non-zero digit in the sixth decimal place, one may safely conclude that a_1 in the left-hand side should be equal to the digit in the third decimal position in the right-hand side, that is $a_1 = 7$. Hence, al-Kashi obtains for the first approximation $x_1 = 0.017$ (the 0th approximation being $x_0 = 0.01$). Then again:

$$0.017a_2a_3\dots = \frac{4}{3}(0.017a_2a_3\dots)^3 + \frac{1}{3} \sin 3^\circ$$

or (subtracting 0.017 from both sides)

$$0.000a_2a_3 \dots = \frac{4}{3}(0.017a_2a_3 \dots)^3 + 0.0004453 \dots$$

He then reasons as before, comparing the fourth decimal digit of the left-hand side to that of the right-hand side, and concluding that $a_2 = 4$, so that the second approximation now reads $x_2 = 0.0174$. In this fashion, al-Kashi computes the value 0.0174524064372835103712 for $\sin 1^\circ$, which is correct up to the first 17 decimal digits. In doing so, he claims that the k th decimal digit of the value of the right-hand side of equation (4) depends only on the values of the first $k - 1$ digits in the decimal expansion of x , and concludes that one can determine the value of $\sin 1^\circ$ with any degree of accuracy.

Let us now examine why al-Kashi's algorithm does indeed produce a value as close to the true value of $\sin 1^\circ$ as one wishes. Denoting $\frac{1}{3} \sin 3^\circ$ by p , $\frac{4}{3}x^3 + p$ by $f(x)$, and $0.01a_1a_2 \dots a_n$ by x_n , we observe that the procedure described above amounts to the iteration of f starting with x_0 :

$$\begin{aligned} x_1 &= f(x_0), \\ x_2 &= f(x_1) = f(f(x_0)), \\ &\dots \\ x_n &= f(f(f(\dots f(x_0)\dots))). \end{aligned} \tag{6}$$

Using a common scientific calculator and starting with $x_0 = 0.01$, one readily calculates $x_4 = 0.017452406437273 \dots$, which provides an approximate value for $\sin 1^\circ$ that is correct up to the first 13 decimal digits.

We have to prove that the iteration equation (6) does converge to $\sin 1^\circ$, so that al-Kashi's algorithm is an effective procedure. To see this, note that equation (4) can be reinterpreted as $r = f(r)$, which means that determining a root of equation (4) is equivalent to finding a fixed point r of the function f . Now, the existence and uniqueness of such a fixed point are guaranteed by the following general theorem [6].

Theorem. *Let f be a continuous function that maps a closed interval $I = [a, b]$ into itself, and for which $|f'(x)| \leq M < 1$ for all x in I . Then, f has a unique fixed point r in I . Moreover, for any x_0 in I , the sequence $\{x_n\}$ obtained by the iterations of f starting with x_0 converges to r , and $|x_n - r| \leq M^n \times |x_0 - r|$ for all $n \geq 1$. (See Figure 3.)*

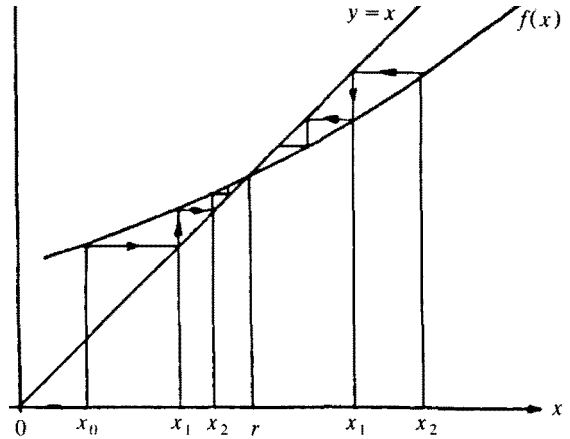


Figure 3.

To apply the above theorem to al-Kashi's cubic $f(x) = \frac{4}{3}x^3 + p$ with $p = \frac{1}{3} \sin 3^\circ$, first note that as p is of the order of 0.01, we can choose the interval I to be, say, $[0.01, 0.02]$. Then, $f'(x) = 4x^2$ is positive on I , so that I is increasing on this interval, and since $f(0.01) > 0.01$ while $f(0.02) < 0.02$, it follows that f maps I into itself. Furthermore, the derivative f' , being increasing on I , assumes its maximum at $x = 0.02$, so that $|f'(x)| \leq 1.6 \times 10^{-3}$. This justifies al-Kashi's claim that each iteration of f yields (at least) one more correct decimal digit, i.e., the error is reduced by a factor at least as small as $\frac{1}{10}$ — the theorem guarantees that, in fact, the error is reduced by at least a factor of 1.6×10^{-3} . Hence, the iteration algorithm does converge to the unique fixed point. The insensitivity of the iteration to the choice of the 0th approximation x_0 , as long as x_0 is chosen in I , becomes evident too. At this point, it is worth mentioning that the condition that f be a "contraction mapping" in the vicinity of the fixed point (i.e., $|f'(x)| < 1$ for all x in I) is absolutely crucial: If this condition is not satisfied, then the iterative procedure described above may generate cyclic or even chaotic sequences, leading away from the fixed point [7].

In conclusion, it is now apparent that besides converging rapidly, al-Kashi's algorithm requires only a few simple operations at each step: raising a number to the third power, an addition, and a division. That al-Kashi did not seem to concern himself with questions of existence and uniqueness should not be held against him. He was primarily interested in devising methods for accurately determining numerical solutions to problems important in astronomy. The foregoing should show how he acquitted himself with deep insight and great elegance.

References

1. A. Aaboe, *Episodes from the Early History of Mathematics*, Random House, New York, 1964, pp. 115–116.
2. A. Aaboe, Al-Kashi's iteration method for the determination of $\sin 1^\circ$, *Scripta Mathematica* 20 (1954) 24–29; see references list.
3. J. L. Berggren, *Episodes in the Mathematics of Medieval Islam*, Springer, New York, 1986.
4. A. A. Birjandi, *Commentary on Ulugh Beg's Astronomical Tables*, Tehran University Library manuscript no. 473 (in Persian).
5. C. B. Boyer, *A History of Mathematics*, 2nd ed., Wiley, New York, 1991.
6. R. L. Burden and J. D. Faires, *Numerical Analysis*, 4th ed., PWS-Kent, Boston, 1988, p. 60.
7. L. O. Cannon and J. Ehrlich, Some pleasures and perils of iteration, *Mathematics Teacher* 86 (1993) 233–239.
8. A. P. Juschkewitsch, *Geschichte der Mathematik im Mittelalter*, Pfalz Verlag, Basel, 1964, pp. 241–242.
9. al-Kashi, *Key to Arithmetic*, Tehran edition, undated (in Arabic); author's translation.
10. V. J. Katz, *A History of Mathematics*, Harper Collins, New York, 1993.

Leonardo of Pisa and his Liber Quadratorum

R.B. McCLENON

American Mathematical Monthly 26 (1919), 1–8

The thirteenth century is a period of great fascination for the historian, whether his chief interest is in political, social, or intellectual movements. During this century great and far-reaching changes were taking place in all lines of human activity. It was the century in which culminated the long struggle between the Papacy and the Empire; it brought the beginnings of civil liberty in England; it saw the building of the great Gothic cathedrals, and the establishment and rapid growth of universities in Paris, Bologna, Naples, Oxford, and many other centers. The crusades had awakened the European peoples out of their lethargy of previous centuries, and had brought them face to face with the more advanced intellectual development of the East. Countless travelers passed back and forth between Italy and Egypt, Asia Minor, Syria, and Bagdad; and not a few adventurous and enterprising spirits dared to penetrate as far as India and China. The name of Marco Polo will occur to everyone, and he is only the most famous among many who in those stirring days truly discovered new worlds.

Among the many valuable gifts which the Orient transmitted to the Occident at this time, undoubtedly the most precious was its scientific knowledge, and in particular the Arabian and Hindu mathematics. The transfer of knowledge and ideas from East to West is one of the most interesting phenomena of this interesting period, and accordingly it is worth while to consider the work of one of the pioneers in this movement.

Leonardo of Pisa, known also as Fibonacci¹, in the last years of the twelfth century made a tour of the East, saw the great markets of Egypt and Asia

Minor, went as far as Syria, and returned through Constantinople and Greece [8]. Unlike most travelers, Leonardo was not content with giving a mere glance at the strange and new sights that met him, but he studied carefully the customs of the people, and especially sought instruction in the arithmetic system that was being found so advantageous by the Oriental merchants. He recognized its superiority over the clumsy Roman numeral system which was used in the West, and accordingly decided to study the Hindu-Arabic system thoroughly and to write a book which should explain to the Italians its use and applications. Thus the result of Leonardo's travels was the monumental *Liber Abaci* (1202), the greatest arithmetic of the middle ages, and the first one to show by examples from every field the great superiority of the Hindu-Arabic numeral system over the Roman system exemplified by Boethius [2]. It is true that Leonardo's *Liber Abaci* was not the first book written in Italy in which the Hindu-Arabic numerals were used and explained [10], but no work had been previously produced which in either the extent or the value of its contents could for a moment be compared with this. Even today it would be thoroughly worthwhile for any teacher of mathematics to become familiar with many portions of this great work. It is valuable reading both on account of the mathematical insight and originality of the author, which constantly awaken our admiration, and also on account of the concrete problems, which often give much interesting and significant information about commercial customs and economic conditions in the early thirteenth century.

Besides the *Liber Abaci*, Leonardo of Pisa wrote an extensive work on geometry, which he called *Practica Geometriae*. This contains a wide variety of interesting theorems, and while it shows no such

¹This is probably a contraction for "Filiorum Bonacci," or possibly for "Filius Bonacci"; that is, "of the family of Bonacci" or "Bonacci's son." See [3].

originality as to enable us to rank Leonardo among the great geometers of history, it is excellently written, and the rigor and elegance of the proofs are deserving of high praise. A good idea of a small portion of the *Practica Geometriae* can be obtained from Archibald's very successful restoration of Euclid's *Divisions of Figures* [1].

The other works of Leonardo of Pisa that are known are *Flos*, a *Letter to Magister Theodorus*, and the *Liber Quadratorum*. These three works are so original and instructive, and show so well the remarkable genius of this brilliant mathematician of the thirteenth century, that it is highly desirable that they be made available in English translation. It is my intention to publish such a translation when conditions are more favorable, but in the meantime a short account of the *Liber Quadratorum* will bring to those whose attention has not yet been called to it some idea of the interesting and valuable character of the book.

The *Liber Quadratorum* is dedicated to the Emperor Frederick II, who throughout his whole career showed a lively and intelligent interest in art and science, and who had taken favorable notice of Leonardo's *Liber Abaci*. In the dedication, dated in 1225, Leonardo relates that he had been presented to the Emperor at court in Pisa, and that Magister Johannes of Palermo had there proposed a problem² as a test of Leonardo's mathematical power. The problem was, to find a square number which when either increased or diminished by 5 should still give a square number as result. Leonardo gave a correct answer, $11\frac{97}{144}$. For $11\frac{97}{144} = (3\frac{5}{12})^2$, $6\frac{97}{144} = (2\frac{7}{12})^2$, and $16\frac{97}{144} = (4\frac{1}{12})^2$. Through considering this problem and others allied to it, Leonardo was led to write the *Liber Quadratorum* [8]. It should be said that this problem had been considered by Arab writers with whose works Leonardo was unquestionably familiar; but his methods are original, and our admiration for them is not diminished by careful study of what had been done by his Arabian predecessors [11].

In the *Liber Quadratorum*, Leonardo has given us a well-arranged, brilliantly written collection of theorems from indeterminate analysis involving equations of the second degree. Many of the theorems themselves are original, and in the case of many others the proofs are so. The usual method of proof employed is to reason upon general numbers, which

Leonardo represents by line segments. He has, it is scarcely necessary to say, no algebraic symbolism, so that each result of a new operation (unless it be a simple addition or subtraction) has to be represented by a new line. But for one who had studied the "geometric algebra" of the Greeks, as Leonardo had, in the form in which the Arabs used it [6], [12], [7], this method offered some of the advantages of our symbolism; and at any rate it is marvelous with what ease Leonardo keeps in his mind the relation between two lines and with what skill he chooses the right road to bring him to the goal he is seeking.

To give some idea of the contents of this remarkable work, there follows a list of the most important results it contains. The numbering of the propositions is not found in the original.

PROPOSITION I. Theorem. Every square number³ can be formed as a sum of successive odd numbers beginning with unity. That is,

$$1 + 3 + 5 + \cdots + (2n - 1) = n^2.$$

PROPOSITION II. Problem. To find two square numbers whose sum is a square number. "I take any odd square I please, ... and find the other from the sum of all the odd numbers from unity up to that odd square itself."⁴ Thus, if $2n + 1$ is a square ($= x^2$), then

$$\begin{aligned} 1 + 3 + 5 + \cdots + (2n - 1) + x^2 &= n^2 + (2n + 1) \\ &= \text{a sum of two squares} = (n + 1)^2. \end{aligned}$$

This is equivalent to Pythagoras's rule for obtaining rational right triangles, as stated by Proclus [9], viz.,

$$\left(\frac{x^2 - 1}{2}\right)^2 + x^2 = \left(\frac{x^2 + 1}{2}\right)^2.$$

For, inasmuch as $2n + 1 = x^2$, we have

$$n = (x^2 - 1)/2 \quad \text{and} \quad n + 1 = (x^2 + 1)/2.$$

PROPOSITION III. Theorem.

$$\left(\frac{n^2}{4} - 1\right)^2 + n^2 = \left(\frac{n^2}{4} + 1\right)^2.$$

This enables us to obtain rational right triangles in which the hypotenuse exceeds one of the legs by 2. It

²In the introduction to *Flos* we are told that two other problems were propounded at the same time.

³Throughout this article, unless otherwise stated, the word "number" is to be understood as meaning "positive integer."

⁴The use of quotation marks indicates a literal translation of Leonardo's words; in other cases the exposition follows his thought without adhering closely to his form of expression.

is attributed by Proclus to Plato [9]. Leonardo also gives the rule in case the hypotenuse is to exceed one leg by 3, and indicates what the result would be if the hypotenuse exceeds one leg by any number whatever.

PROPOSITION IV. Theorem. “Any square exceeds the square which immediately precedes it by the amount of the sum of their roots.” That is,

$$n^2 - (n - 1)^2 = n + (n - 1).$$

It follows from this that when the sum of two consecutive numbers is a square number, then the square of the greater will equal the sum of two squares. For, if $n + (n - 1) = u^2$, then $n^2 - (n - 1)^2 = u^2$ or $n^2 = u^2 + (n - 1)^2$.

PROPOSITION V. Problem. Given $a^2 + b^2 = c^2$, to find two integral or fractional numbers x, y , such that $x^2 + y^2 = c^2$.

Solution: By Proposition II or Proposition III, find two other numbers m and n such that $m^2 + n^2 = q^2$. If $q^2 \neq c^2$, multiply the preceding equation by c^2/q^2 , obtaining

$$\left(\frac{c}{q} \cdot m\right)^2 + \left(\frac{c}{q} \cdot n\right)^2 = c^2$$

so that $x = c/q \cdot m, y = c/q \cdot n$ is a solution.

PROPOSITION VI. Theorem. “If four numbers not in proportion are given, the first being less than the second, and the third less than the fourth, and if the sum of the squares of the first and second is multiplied by the sum of the squares of the third and fourth, there will result a number which will be equal in two ways to the sum of two square numbers.” That is,

$$\begin{aligned} (a^2 + b^2)(c^2 + d^2) &= (ac + bd)^2 + (ad - bc)^2 \\ &= (ad + bc)^2 + (ac - bd)^2. \end{aligned}$$

This very important theorem should be called Leonardo’s Theorem, for it is not found definitely stated, to say nothing of being proved, in any earlier work. Leonardo considers also the case where a, b, c , and d are in proportion, and shows that then $(a^2 + b^2)(c^2 + d^2)$ is equal to a square and the sum of two squares. This gives him still another way of finding rational right triangles.⁵

⁵For instance, letting $a = 6, b = 4, c = 3, d = 2$, we have $(36 + 16)(9 + 4) = 676 = (6 \cdot 3 + 4 \cdot 2)^2 = (6 \cdot 2 + 4 \cdot 3)^2 + (6 \cdot 3 - 4 \cdot 2)^2 = 26^2 = 24^2 + 10^2$.

PROPOSITION VII. Theorem.

$$(x^2 - y^2)^2 + (2xy)^2 = (x^2 + y^2)^2$$

This is Euclid’s general solution of the problem of finding rational right triangles [6]; Leonardo proves this very simply as a corollary of Proposition VI.

PROPOSITION VIII. Problem. “To find two numbers the sum of whose squares is a number, not a square, formed from the addition of two given squares.” That is, to find x and y such that $x^2 + y^2 = a^2 + b^2$. Choose any two numbers c and d , such that $c^2 + d^2$ is a square, and write $(a^2 + b^2)(c^2 + d^2)$ as a sum of two squares, let us say $p^2 + q^2$; this we can do by Proposition VI. Construct the right triangle whose legs are p and q ; then the similar triangle whose hypotenuse is equal to $\sqrt{c^2 + d^2}$ will have as its legs the two required numbers x and y .

PROPOSITION IX. Theorem.

$$6(1^2 + 2^2 + 3^2 + \dots + n^2) = n(n + 1)(2n + 1).$$

The proof of this is strikingly original, and proceeds from the identity

$$n(n + 1)(2n + 1) = n(n - 1)(2n - 1) + 6n^2.$$

Hence

$$\begin{aligned} n(n - 1)(2n - 1) &= (n - 1)(n - 2)(2n - 3) \\ &\quad + 6(n - 1)^2, \end{aligned}$$

...

$$(2)(3)(2 + 3) = (1)(2)(1 + 2) + 6(2)^2,$$

$$(1)(2)(1 + 2) = 6(1)^2.$$

It follows by addition that

$$\begin{aligned} n(n + 1)(2n + 1) &= 6(1^2 + 2^2 + 3^2 \\ &\quad + \dots + (n - 1)^2 + n^2). \end{aligned}$$

PROPOSITION X. Theorem.

$$\begin{aligned} 12[1^2 + 3^2 + 5^2 + \dots + (2n - 1)^2] \\ = (2n - 1)(2n + 1)4n. \end{aligned}$$

Leonardo gives a proof very similar to that of Proposition IX.

PROPOSITION XI. Theorem.

$$12[2^2 + 4^2 + 6^2 + \dots + (2n)^2] = 2n(2n + 2)(4n + 2),$$

and likewise

$$18[3^2 + 6^2 + 9^2 + \dots + (3n)^2] = 3n(3n + 3)(6n + 3),$$

and

$$24[4^2 + 8^2 + 12^2 + \cdots + (4n)^2] = 4n(4n+4)(8n+4),$$

and in general

$$6a[a^2 + (2a)^2 + (3a)^2 + \cdots + (na)^2] \\ = na(na + a)(2na + a).$$

Here Leonardo has almost discovered the general result

$$a^2 + (a + d)^2 + (a + 2d)^2 + \cdots + [a + (n - 1)d]^2 \\ = \frac{6na^2 + 6n(n - 1)ad + n(n - 1)(2n - 1)d^2}{6}.$$

His method needed no change at all, in fact.

PROPOSITION XII. Theorem. If $x + y$ is even, $xy(x + y)(x - y)$ is divisible by 24; and in any case $4xy(x + y)(x - y)$ is divisible by 24. A number of this form is called by Leonardo a *congruum*, and he proceeds to show that it furnishes the solution to a problem proposed by Johannes of Palermo.

PROPOSITION XIII. Problem. "To find a number which, being added to, or subtracted from, a square number, leaves in either case a square number." Leonardo's solution of this, the problem which had stimulated him to write the *Liber Quadratorum*, is so very ingenious and original that it is a matter of regret that its length prevents its inclusion here. It is not too much to say that this is the finest piece of reasoning in number theory of which we have any record, before the time of Fermat. Leonardo obtains his solution by establishing the identities

$$(x^2 + y^2)^2 - 4xy(x^2 - y^2) = (y^2 + 2xy - x^2)^2$$

and

$$(x^2 + y^2)^2 + 4xy(x^2 - y^2) = (x^2 + 2xy + y^2)^2.$$

PROPOSITION XIV. Problem. To find a number of the form $4xy(x + y)(x - y)$ which is divisible by 5, the quotient being a square. Take $x = 5$, and y equal to a square such that $x + y$ and $x - y$ are also squares. The least possible value for y is 4, in which case

$$4xy(x + y)(x - y) = (4)(5)(4)(9)(1) = 720.$$

PROPOSITION XV. Problem. "To find a square number which, being increased or diminished by 5,

gives a square number. Let a congruum be taken whose fifth part is a square, such as 720, whose fifth part is 144; divide by this the squares congruent to 720,⁶ the first of which is 961, the second 1681, and the third 2401. The root of the first square is 31, of the second is 41, and of the third is 49. Thus there results for the first square $6\frac{97}{144}$, whose root is $2\frac{7}{12}$, which results from the division of 31 by the root of 144, that is, by 12; and for the second, that is, for the required square, there will result $11\frac{97}{144}$, whose root is $3\frac{5}{12}$, which results from the division of 41 by 12; and for the last square there will result $16\frac{97}{144}$, whose root is $4\frac{1}{12}$."

PROPOSITION XVI. Theorem. When $x > y$, $(x + y)/(x - y) \neq x/y$. It follows that $x(x - y)$ is not equal to $y(x + y)$, and "from this," Leonardo says, "it may be shown that no square number can be a congruum." For if $xy(x + y)(x - y)$ could be a square, either $x(x - y)$ must be equal to $y(x + y)$, which this proposition proves to be impossible, or else the four factors must severally be squares, which is also impossible. Leonardo to be sure overlooked the necessity of proving this last assertion, which remained unproved until the time of Fermat [4], [5].

PROPOSITION XVII. Problem. To solve in rational numbers the pair of equations

$$x^2 + x = u^2, \quad x^2 - x = v^2.$$

The solution is obtained by means of any set of three squares in arithmetic progression, that is, by means of Proposition XIII. Let us take x_1^2 , x_2^2 , and x_3^2 for the three squares, and let the common difference, that is, the congruum, be d . Leonardo says that the solution of the problem is obtained by giving x the value x_2^2/d . For then⁷

$$x^2 + x = \frac{x_2^4}{d^2} + \frac{x_2^2}{d} = \frac{x_2^2(x_2^2 + d)}{d^2} = \frac{x_2^2 x_3^2}{d^2};$$

and

$$x^2 - x = \frac{x_2^4}{d^2} - \frac{x_2^2}{d} = \frac{x_2^2(x_2^2 - d)}{d^2} = \frac{x_2^2 x_1^2}{d^2}.$$

⁶That is, the three squares in arithmetic progression, whose common difference is the congruum 720. They are obtained by Proposition XIII, thus: Taking $x = 5$ and $y = 4$, $y^2 + 2xy - x^2 = 31$, the root of the first square; $x^2 + y^2 = 41$, the root of the second square; and $x^2 + 2xy + y^2 = 49$, the root of the third square.

⁷The simplest numerical example would be $x_1^2 = 1$, $x_2^2 = 25$, $x_3^2 = 49$, and this is the illustration given by Leonardo. It leads to $x = 25/24$, from which we have $x^2 + x = 1225/576 = (35/24)^2$ and $x^2 - x = 25/576 = (5/24)^2$.

PROPOSITION XVIII. Problem. To solve in rational numbers the pair of equations

$$x^2 + 2x = u^2, \quad x^2 - 2x = v^2.$$

The method is similar to that in Proposition XVII, the value of x being found to be $2x_2^2/d$. Leonardo adds, “You will understand how the result can be obtained in the same way if three or more times the root is to be added or subtracted.”

PROPOSITION XIX. Problem. To solve (in integers) the pair of equations

$$x^2 + y^2 = u^2, \quad x^2 + y^2 + z^2 = v^2.$$

Take for x and y any two numbers that are prime to each other and such that the sum of their squares is a square, let us say u^2 . Adding all the odd numbers from unity to $u^2 - 2$,⁸ the result is $((u^2 - 1)/2)^2$. Now

$$\left(\frac{u^2 - 1}{2}\right)^2 + u^2 = \left(\frac{u^2 + 1}{2}\right)^2.$$

Thus

$$z^2 = \left(\frac{u^2 - 1}{2}\right)^2, \quad \text{and} \quad v^2 = \left(\frac{u^2 + 1}{2}\right)^2.$$

PROPOSITION XX. Problem. To solve in rational numbers the set of equations

$$\begin{aligned} x + y + z + x^2 &= u^2, \\ x + y + z + x^2 + y^2 &= v^2, \\ x + y + z + x^2 + y^2 + z^2 &= w^2. \end{aligned}$$

By an extension of the method used in Proposition XIX Leonardo obtains the results $x = 3\frac{1}{5}$, $y = 9\frac{3}{5}$, $z = 28\frac{4}{5}$. He even goes farther and obtains the integral solutions $x = 35$, $y = 144$, $z = 360$. He continues, “And not only can three numbers be found in many ways by this method but also four can be found by means of four square numbers, two of which in order, or three, or all four added together make a square number . . . I found these four numbers, the first of which is 1295, the second $4566\frac{6}{7}$, the third $11417\frac{1}{7}$, and the fourth 79920.” In the midst of the

explanation of how these values were obtained, the manuscript of the *Liber Quadratorum* breaks off abruptly. It is probable, however, that the original work included little more than what the one known manuscript gives. At all events, considering both the originality and power of his methods, and the importance of his results, we are abundantly justified in ranking Leonardo of Pisa as the greatest genius in the field of number theory who appeared between the time of Diophantus and that of Fermat.

Bibliography

1. Archibald, *Euclid's book on Divisions of Figures, with a Restoration based on Woepcke's Text and on the Practica Geometricae of Leonardo Pisano*, Cambridge, England, 1915.
2. *Boethius*, ed. Friedlin, Leipzig, 1867.
3. Boncompagni, *Della Vita e delle Opere di Leonardo Pisano, matematico del secolo decimoterzo*, Rome, 1852.
4. Fermat, *Oeuvres*, Paris, 1891.
5. Heath, T. L., *Diophantus of Alexandria*, Cambridge, 1910.
6. Heath, T. L., *The Thirteen Books of Euclid's Elements*, Cambridge, 1908.
7. Karpinski, L. C., *Robert of Chester's Latin translation of the Algebra of Al-Khowarizmi*, New York, 1915.
8. Leonardo of Pisa, *Scritti di Leonardo Pisano*, 2 vols., Rome, 1857–1861.
9. *Proclus*, ed. Friedlein, Leipzig, 1873.
10. Smith and Karpinski, *The Hindu-Arabic Numerals*, Boston and London, 1911.
11. Woepcke, *Recherches sur plusieurs ouvrages de Leonard de Pise, et sur les rapports qui existent entre ces ouvrages et les travaux mathématiques des Arabes*, Rome, 1859.
12. Zeuthen, H. G., *Geschichte der Mathematik im Alterum und Mittelalter*, Copenhagen, 1896.

⁸Here u^2 is odd, because it is the sum of the squares of two numbers x and y which are prime to each other. It is not possible that both x and y are odd, since $(2m + 1)^2 + (2n + 1)^2 = 4m^2 + 4m + 4n^2 + 4n + 2$, and this is divisible by 2 but not by 4, and hence can not be a square. Thus, of the numbers x and y , one must be even and the other odd, hence $x^2 + y^2$ is odd.

The Algorists vs. the Abacists: An Ancient Controversy on the Use of Calculators

BARBARA E. REYNOLDS

College Mathematics Journal 24 (1993), 218–223

In 1299 the bankers of Florence were forbidden to use Arabic numerals and were obliged instead to use Roman numerals. And in 1348 the University of Padua directed that a list of books for sale should have the prices marked “non per cifras, sed per literas clara” (not by figures, but by clear letters). [1], [11], [15]

Our “modern” decimal system of notation actually comes to us from ancient India. Some of the symbols in use today were used as early as the third century B.C. (The zero, however, did not appear until much later — about A.D. 376.) The Arabs carried these numerals into Western Europe at the time of the Moorish invasions about A.D. 750. Gerbert, who became Pope Sylvester II toward the end of the tenth century, is the first European scholar who is definitely known to have taught using the Hindu-Arabic numeration system. Yet, three and four hundred years later we find these numerals being outlawed! Hindu-Arabic numerals seem so much more convenient to use than Roman numerals, especially for representing large numbers in a small space, that we might wonder why this system of notation was not readily adopted as soon as it was known. [3], [7], [13]

Since the Greeks of the sixth and fifth centuries B.C. are known to have traveled throughout the ancient world and would certainly have come into contact with positional systems of numeration such as that used by the Babylonians, it seems strange that they did not generally recognize and adopt a numeration system that was more efficient for computation than their own non-positional system. Various conjectures have been offered: that they were more interested in properties of the numbers themselves than in theories of computation, or that they preferred to

guard knowledge of computation from common people (and thus protect the hold the educated elite held over the lower classes). In either case, the ancient Greeks, who contributed much to the development of other areas of mathematical thought, did little to advance the use of a positional numeration system, and thus little to advance the science of computation. [9]

The numerals used by the ancient Romans are familiar to us, as they are still used for such things as numbering chapters in a book or representing the date on a cornerstone. Suppose Roman numerals were the only numerals we had, and that all computations had to be done using them. The notation is simple enough for small numbers when the calculations can be easily done mentally, but it quickly becomes cumbersome as the numbers get large.

How did the ancient Romans do their calculations? For instance, how would they have found the sum CCVIII + DCXVII? Perhaps they first wrote down all the symbols that appeared in the problem, thus: DCCCXVIII, and then regrouped these as: DCCCXXV. (See Figure 1.) However, there is archeological evidence to suggest that normally they would have used a calculating board or abacus.

Contemporary writers in ancient Greece and Rome made frequent reference to the use of pebbles and

$$\begin{array}{r}
 \text{CCVIII} \\
 + \text{DCXVII} \\
 \hline
 \text{D C X V III} \\
 \text{CC V III} \\
 \hline
 \text{D CCC XX V}
 \end{array}$$

Figure 1.



Figure 2. *The abacist versus the algorist.* (From Gregor Reisch, *Margarita Philosophica*, Strassbourg, 1504.)

other counters for reckoning. Although no wooden boards have survived from those times, several stone or marble Greek abaci have been found. At many Roman sites, archeologists have found piles of small smooth rounded stones which look like they may have been made by dropping a hot molten substance from several feet above a flat surface. These stones would have been a convenient size to use in calculating on an abacus. A few bead-frame calculators, small enough to be held in one hand, have been found. Perhaps these were the portable pocket calculators of that day. [11]

Now how would that same addition problem have appeared if the Romans did the actual computation on the abacus? Unlike paper and pencil algorithms, all computations on the abacus are dynamic processes. The first number is represented on the board

by appropriately placed stones or counters. To add the second number to this, more stones are pushed into place and the regrouping is done quickly — almost “automatically” — so that when the computation is finished, the only number appearing on the board is the final result.

In Figure 3, notice that V is represented by a stone in the space between the I and the X, as if it were an intermediate grouping. Similarly D is represented by a stone in the space between C and M. The five I-stones are replaced by one stone in the V-space, and two V-stones are replaced by one stone on the X-line. The final result would appear as in Figure 3b. This could easily be recorded as DCCCXXV by glancing at the stones. In the notation which we use today, those seven stones would be represented by just three symbols as 825.

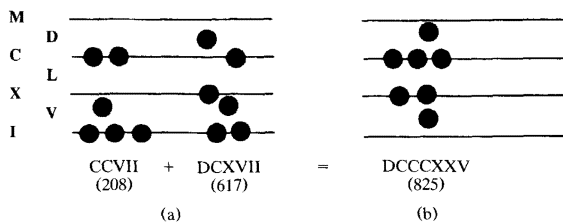


Figure 3.

The examples of abacus representations and the corresponding Roman and Arabic notations in Figure 4 suggest that Roman numerals are easier to use in recording calculations that were done on the counting board. An illiterate public may well have been suspicious of this new notation which sometimes used one symbol to represent two, three, or even four stones and — worse — sometimes used a symbol where there was no stone at all! Although the use of the zero symbol was sometimes understood in medieval times, it was, for the most part, a confusing concept which was not needed in the use of the counter board or in the writing of Roman figures. If the result was recorded in Roman numerals, you would have exactly one symbol for each stone on the counter board. Pen-reckoning, as it was called, and even the figures themselves were treated with suspicion. [7], [11]

The Italian laws referred to at the beginning of this article hint of some kind of controversy. Generally laws are not enacted unless the state feels the need to protect one group of individuals from another.

Although the Hindu system of numeration had been rejected by some, Italian merchants of the twelfth century recognized its superiority for computational purposes. These merchants became noted for their knowledge of arithmetic operations and developed methods of double-entry bookkeeping. However, the forms of the Hindu numerals were not fixed, and the variety of forms gave rise to ambiguity and fraud. (Human nature hasn't changed appreciably since the twelfth century!) Outside of Italy, most European merchants kept accounts in Roman numerals until at least 1550 (and most colleges and monasteries until 1650!). [1], [4]

The struggle between the algorists, as the advocates of pen-reckoning were called, and the abacists continued into the sixteenth century. In manuscripts from the twelfth century, there are striking differences between these two groups. Algorists calculate with a zero, do not employ the abacus, teach extraction of roots, and use Babylonian sexagesimal

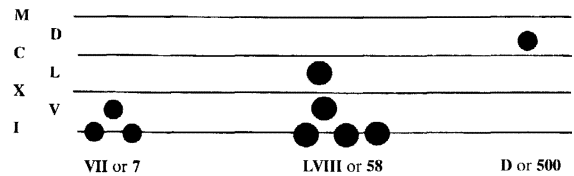


Figure 4.

fractions; the abacists, on the other hand, make no reference to Hindu-decimal notations, use abacus methods (which make extraction of roots almost unthinkable), and use Roman duodecimal fractions. [4], [8]

In Germany, France, and England, Hindu numerals were scarcely used before the mid-fifteenth century. The use of the abacus seems to have hung on well into the seventeenth and eighteenth centuries. Even after people finally began to trust the Arabic numerals, they still preferred to use abacus methods to do their calculations. Evidence of the widespread popular use of abacus methods can be found in arithmetic books published in the sixteenth through the eighteenth centuries. [4], [10], [11]

Among the most popular English-language textbooks were those written by Robert Recorde. In his *Ground of Artes Teaching Works and Practice of Arithmetik*, first published in 1542, he included a chapter showing the use of abacus methods for doing calculations using Hindu-Arabic figures. This chapter was retained in subsequent editions for more than a hundred years! Numerous other arithmetics published in the sixteenth through the eighteenth centuries in French, Spanish, German, and even Latin describe abacus methods for solving arithmetic problems. [10], [11]

A few tables have survived from the late Middle Ages with lines carved into their tops and letters or figures carved into them indicating that they may have been used as counting tables. And there are a large number of metal counters for use on these tables. (These counters are called *Rechen-pfennig* — literally, “reckoning penny” — in German, and *jeton* in French, from *jeter*, to throw or to cast.) They were minted in great quantities from the twelfth through the eighteenth centuries. [11]

Rechen-pfennig were apparently made to order for the customer and the variety of design is marvelous. Many bear crests or seals of various kinds, or heads of reigning monarchs in England and France. Many carry a proverb or other common saying, such as: HEUT ROT MORGEN TODT (Here today, gone to-

morrow), GOTES SEGEN MACHET REICH (The blessing of God makes it rich), DAS WORT GOTES BLEIBT EWIG (The word of God endures forever). Some of the most interesting counters — and, from the point of view of the history of arithmetic, the most valuable — are those that show a merchant sitting at a counting board. Others show an abacus on the obverse and a problem worked in Arabic numeral on the reverse, indicating that the two methods of computation were in use at the same time. [11]

The thought process involved in working on the counter board is different from that involved in pen-reckoning. If I am solving a problem like $400 - 382$, I begin by saying, “Ten minus two is eight.” This is a fact that I have memorized and just use without picturing in my mind a physical model that represents this fact. But if I make purchases at the drug-store amounting to \$3.82 and give the clerk four one-dollar bills, I expect a dime, a nickel, and three pennies in change. This process of seeing a simple subtraction in concrete terms and of grouping the result in convenient units — nickels, dimes, quarters — is more like the thought process involved in working on an abacus.

Today the abacus is commonly thought of as a “Chinese calculator.” But it did not appear in China until rather late. In fact, the earliest mention of the abacus in Chinese literature does not appear until the twelfth century A.D.! [11] In places throughout the world where the abacus is presently being used — Japan, China, Russia, and other countries in the Near and Far East — it generally takes the form of beads sliding on fixed bamboo rods. There are a fixed number of beads on each rod and there is no possibility of “carrying” a bead from one rod to the next. Usually the beads are arranged on each rod with the unit-beads below a calculating bar and the five-beads above the bar, as in Figure 5. When beads are pushed against this bar, they are considered in the calculation; when they are pushed away from the bar, they are ignored. So the beads on the abacus pictured in Figure 5 represent the value 71,536.

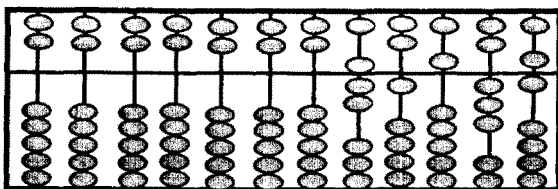


Figure 5.

The usual form of the Chinese abacus or *suan-pan* has five unit-beads and two five-beads on each rod. So it would be possible to represent a value of up to 15 on each rod. Then in simplifying a result, two five-beads on one rod are pushed away from the bar while a unit-bead is pushed toward the bar on the rod immediately to the left. Five unit-beads at the bar are replaced in a similar manner by one five-bead on the same rod. Thus, when the result is completely simplified, the value in ordinary decimal notation can be read from the abacus by simply reading the value represented on each rod from left to right.

The Japanese have developed a form of the abacus which allows faster operation by a skilled person. The beads are smaller and each rod is shorter. (Thus each bead is moved through a shorter distance than on the Chinese model.) Also, the Japanese abacus or *soroban* has only four unit-beads and one five-bead on each rod. So it is possible to count up only to 9 on each rod. In a manner very similar to ordinary decimal notation, the regrouping must be done constantly throughout the process of calculation. The final answer is easily read without any need to simplify the result.

A little reflection on the different numbers of beads on each rod of the Chinese and Japanese abaci gives us a method for doing calculations in various bases. In general, in order to do calculations in base n , we would need an abacus on which we could count to at least $n - 1$ on each rod. An ordinary Chinese abacus can be used to work problems in base 16, or hexadecimal notation. Similarly, if we ignore the five-beads and consider only the five unit-beads, we could do calculations in base 6. And by ignoring the five-beads on the Japanese *soroban*, we can calculate in base 5. (In these lower bases, base 5 or base 6, it really isn't necessary to have an intermediate grouping as we can quickly “see” three, four, or five beads.)

The thought processes involved in using the abacus are very closely related to the process of counting, and are surprisingly similar to the arithmetic of making change. For more than two thousand years, the abacus has been used throughout the world to do base-10 arithmetic. Yet the early historic roots of the abacus are older than the common acceptance of decimal notation. Roman numerals, awkward for “pen-reckoning,” seem to be a natural notation for recording the results of calculations done on an abacus. The abacus developed as a mechanical device which operates on basic ideas about counting. In contemporary America we no longer use counting

boards; however, we are dependent upon other kinds of calculating devices. Even modern electronic computers operate on basic ideas about counting — although, these usually count in base 2 at speeds so fast that we easily overlook the underlying counting process.

References

1. W. W. Rouse Ball, *A Short Account of the History of Mathematics*, Dover, New York, 1960. [An unabridged and unaltered republication of the author's fourth edition which first appeared in 1908.]
2. David Bergamini, *Mathematics*, Life Science Library, New York, 1972.
3. J. Bronowski, *The Ascent of Man*, Little Brown, Boston, 1973.
4. Florian Cajori, *A History of Elementary Mathematics*, Macmillan, New York, 1897.
5. Florian Cajori, *A History of Mathematical Notations, Vol. 1: Notations in Elementary Mathematics*, Open Court, LaSalle, IL, 1928.
6. Florian Cajori, *A History of Mathematics*, 2nd ed., Macmillan, New York, 1922.
7. Tobias Dantzig, *Number: The Language of Science*, 4th ed., Free Press, New York, 1954.
8. Howard Eves, *An Introduction to the History of Mathematics*, revised ed., Holt Rinehart Winston, New York, 1964.
9. Bernard H. Grundlach, The history of numbers and numerals, *Historical Topics for the Mathematics Classroom*, Thirty-first Yearbook of the National Council of Teachers of Mathematics, Washington DC, 1969.
10. Louis Charles Karpinski, *The History of Arithmetic*, Russell & Russell, New York, 1965.
11. J. M. Pullan, *The History of the Abacus*, Praeger, New York, 1969.
12. David Eugene Smith, *Number Stories of Long Ago*, National Council of Teachers of Mathematics, Washington, DC, 1919.
13. David Eugene Smith, *History of Mathematics*, 2 vols., Dover, New York, 1958.
14. David Eugene Smith and Jekuthiel Ginsberg, *Numbers and Numerals*, National Council of Teachers of Mathematics, Washington, DC, 1971.
15. Dirk J. Struik, *A Concise History of Mathematics*, 3rd revised ed., Dover, New York, 1967.

Sidelights on the Cardan-Tartaglia Controversy

MARTIN A. NORDGAARD

National Mathematics Magazine 13 (1937–38), 327–346

1

There is quite a difference in the frame of mind which comes with the answer to a problem only vaguely defined and lying in an uncharted field, like the invention of the differential calculus, or with a discovery that comes undivined like a flash of lightning from some human mind, like the invention of logarithms,—and the reaction that greets the answer to a problem posed to the world for centuries when that answer arrives, two thousand years in the coming.

The solution of the cubic had presented itself to the human mind as an intellectual problem already in the fifth century B.C.; it became a scientific need in Archimedes' calculation on floating bodies in the third century B.C.; it confronted the Arab astronomers in the Middle Ages. And now it was solved! The first of "the three unsolved problems of antiquity" to be solved.

It produced a great impression. How great, one can gauge from the fact that all respectable texts on algebra for the next 200 years gave long chapters and discussions to the cubic equation. The influence of the discovery must be gauged not only by its mathematical fruitfulness, which after all did not prove to be so very great, but by the stimulus it gave to study, the courage it gave the human mind to soar into the unknown and "make the impossible possible".

The main events leading up to the discovery of a general solution of the cubic equation and the ensuing controversy are given in the various histories of mathematics. But there are illuminating sidelights in this unique controversy—documentary, anecdotal, biographical—which do not lend themselves to recording in a well-balanced history of mathematics but which are of absorbing interest to the members of the guild of mathematicians. There are the many

source materials, for one thing; from some of these we shall quote extracts. There is the language and symbolism, or lack of it, of the algebra prior to Vieta, Stevin, and Descartes. And then there is the exposition of the status of algebraic theory before the monumental works of Cardan and Tartaglia.

The 16th-century custom of scientific "duels" and public disputations were a joint inheritance from the philosophical disputations of the Schoolmen and the tournaments of the knights. A chief canon of combat was that no one should propose a question or problem that he himself could not solve. The outward forms were modeled somewhat after the contests of arms—challenge, response, witnesses, judges, keeper of the stakes, etc.

Public challenges were given, not only for acquiring glory and prestige, but also for making a living. The vanquished, honor lost, had no more pupils; while the victor, heralded and feted, would be called to various cities to teach and lecture. Consequently, many inventors guarded their secrets. There must have been many discoveries lost to the world due to this custom. Tartaglia himself died while still writing on his algebra and before reaching his contemplated climax on his solution of the cubic; and except for the premature publication of it by Cardan and Tartaglia's accusation in the *Quesiti* his solution might have died with him.

2

The *Dramatis Personae* of the celebrated controversy were five: Zuanne de Tonini da Coi, Antonio Maria Fior, Girolamo Cardano, Nicolo Tartaglia, and Ludovico Ferrari. The time: 1530 to 1548. Place: Pavia, Padua, Bologna, Milano, Brescia, Venice, the centers of art and learning in Renaissance Italy.

The first two were minor characters and little is known about them outside their connection with this controversy; they were messengers, links, as it were, to bring about action between the other three. *Zuanne da Coi* (sometimes called Giovanni dal Colle) was a teacher in Brescia interested in mathematics from the standpoint of problem solving. *Antonio Maria Fior* (sometimes written Floridus and Del Fiore) flitted about from place to place, causing battle and disturbance; but History will thank him for it. He was an arithmetician, having according to reports no theoretical knowledge in algebra. He had been a pupil of Scipio Ferro, of whom more later.

Nicolo Tartaglia was born at Brescia in 1506, died at Venice in 1557. He came from a very poor family, was left fatherless at the age of six, and had only two weeks of formal schooling; but by self education his powerful mind mastered both the classics and the then known mathematics. He taught mathematics in Verona, Vicenza, Brescia, and, from 1534 or 1535 until his death in 1557, in Venice. His principal mathematical works are: *Nova Scienza* (1557), where he is the first one to discuss the problems of gunnery and fortification mathematically; (2) *Questiti ed invenzioni diverse* (1546), in nine books, of which the last one deals with algebra; (3) *General Trattato di numerie misure* in two volumes (the first published in 1556, the second in 1560) including an arithmetic, a treatise on numbers, and his work on algebra.

Girolamo Cardano (Hieronymus Cardanus or Jerome Cardan) was born at Pavia in 1501, died in Rome in 1576. He received a good university education in Pavia and Padua, having equal zest for medicine and mathematics. Between 1524 and 1550 he taught and practiced medicine, much of the time in Milano; in the same period he studied mathematics assiduously and published many important works. In 1562 he became a university professor at Bologna and in 1570 he moved to Rome to become astrologer to the Pope. He wrote voluminously on many subjects, but in mathematics we mention these: (1) *Practicae Arithmeticae* (1539); (2) *De Regula Aliza* (1540); (3) *Ars Magna* (1545), the first systematic work in algebra up to that time, a text that helped to clarify the principles of algebra and lift the subject out of mere equational problem solving into a theory of equations.

Ludovico Ferrari was born at Bologna in 1522 and died there in 1565. His parents were poor, and he came to Cardan's house as an errand boy. He was later allowed to listen to his master's lectures, and

before long became Cardan's most brilliant pupil. For all his moral lapses and irascible temper, Ferrari was ever loyal to his protector; in fact, he looked upon himself as owing his very being to Cardan, designating himself as "suo creato". As far as we know he never published anything independently. What he discovered he let Cardan incorporate into the *Ars Magna*. One of his discoveries was a general solution of the biquadratic equation. For he was able, by using the solution of a cubic already discovered by Tartaglia, to solve the question proposed by Da Coi, namely

$$x^4 + 6x^2 + 36 = 60x,$$

succeeding where both Tartaglia and Cardan had failed. Ferrari was only twenty-three years old when the *Ars Magna* was published.

In reading works on algebra from this period the reader must learn to divorce from his consciousness many of the ideas and forms he has associated with algebra. He will remember that in 1500 there were no imaginary numbers; they did at times make unwelcomed appearances but were not legitimized. Negative numbers did not have operational status and $x^3 = px + q$ had a different solution from that of $x^3 + px = q$, for instance. The symbols +, -, =, in our sense, did not exist, and our words "plus" and "minus" were not conventionalized. The unknown was variously called *thing*, *side*, *cos*, *res*. Thus Tartaglia's equation ("capitolo") $x^3 + 3x^2 = 5$ was "a cube and three censi are equal to five".

Besides Cardan's *Ars Magna* and Tartaglia's *Questiti* and *General Trattato* we have as source material the six *Cartelli* (letters of challenge of Ferrari) and the six *Risposti* (responses) of Tartaglia. These were sent as printed pamphlets to the mathematicians of Italy.

Being literature of a day it is a wonder that all the twelve bulletins have come down to us. As one might expect, they have their own exciting history. In 1844 Prof. Silvestro Gherhardi owned a volume containing the six *Cartelli* of Ferrari and the first five *Risposti* of Tartaglia. In 1848, after a four years' search in all the libraries and old book-stores of the various cities of Italy, Gherhardi finally laid hands on the missing 6th *Risposta* in an old book shop in Bologna; it is the only copy of this *Risposta* found so far. Previously all that was known of Tartaglia's 6th letter were citations from Bombelli (1572) and writers living later than Tartaglia by 200 years. In 1858 Gherhardi, meeting with political vicissitudes and exile and in need of money, sold his copy to Libri

of London, it being “clipped” on to the other eleven. But first Gherardi was permitted to make an exact copy of the letters “by the hand of Benaducci di Foligno”. And that was fortunate; for the copy sent to Libri was lost. So it has been by a slender thread that the last of the twelve letters has reached us. In 1876 the twelve letters were “collected, autographed, and published” by Enrico Giordani in a limited edition of 212 copies under the title *I sei cartelli di matematica disfida di Ludovico Ferrari con sei controcartelli in risposta di Nicolo Tartaglia*.

An additional word concerning Tartaglia’s *Quesiti ed invenzioni diverse*. It consists of short, sprightly accounts in dialogue form of problems he discussed with or solved for various people, the first *Quesiti* dated 1521, the last, 1541. Quoting conversations and letters, citing dates, places, and names of interlocutors, many of whom were still living, the book has strong documentary secureness. It is charmingly written, besides. The last of the nine books, comprising 42 *quesiti*, deals with algebra.

3

The first to give a general solution for a cubic equation was neither Cardan nor Tartaglia. That honor belongs to Scipio del Ferro, professor of mathematics at Bologna.

As late as 1494 Luca Pacioli in his authority-carrying *Summa* had set forth these types of equations as not yet being solved:

$$\begin{aligned} n &= ax + bx^3 \\ n &= ax^2 + bx^3 \\ n &= ax^3 + bx^4. \end{aligned}$$

And he intimated that their solutions might not be possible. However, the first of these was solved by Ferro of Bologna.

About all we know of Scipio Ferro is that he was born at Bologna about 1465 and died there in 1526, that he was professor at the University of Bologna from 1496 to 1526, that he had a general solution for $x^3 + px = q$, and that he confided his method to his pupil Antonio Maria Fior. We do not know whether he had derived it himself or found it in an Arab work; whether it was an empirical formula or was the product of a demonstration. What writings he left must have come into the hands of his son-in-law Annibale della Nave, who succeeded him in his professorship (1526–1560). But no such writings are extant. Both Cardan and Tartaglia refer to the

solution Fior received from Ferro, Tartaglia placing it about 1506, Cardan placing it at about 1514. It was probably even later.

4

Curiously enough, the one who seems to have set the wheels in motion for the final onslaught on the cubic equation was a man of meagre mathematical attainment but of much physical mobility. It was Zuanne de Tonini da Coi. Teaching in Brescia he had, of course, heard of the work of Nicolo of Brescia, now of Verona. In 1530, as one Brixillite to another, with more courage than prudence he proposed to Tartaglia two problems which reduced to solving the equations

$$x^3 + 3x^2 = 5 \quad \text{and} \quad x^3 + 6x^2 + 8x = 1000.$$

This, the opening chapter in the history of the exciting discovery, is described by Tartaglia in his *Quesiti*, namely in Quesito XIV. There for the first time we learn that Tartaglia (at this time only 24 years of age) had been dabbling with the cubic.

It will give the reader a little of the flavor of the period and give him a peek into one of the interesting books of mathematics to read Tartaglia’s own first reference to the attack on the cubic equation.

“QUESITO XIV, which was proposed to me at Verona by one Maestro Zuanne de Tonini da Coi, who has a school in Brescia, and was brought to me by Messer Antonio da Cellatico in the year 1530.

Maestro Zuanne— Find a number which multiplied by its root increased by three equals five. Similarly find three numbers such that the second is greater by two than the first and the third is greater by two than the second and where the product of the three is 1000.

N.— M. Zuanne, you have sent me these two questions of yours as something impossible to solve or at least as not being known by me; because, proceeding by algebra, the first leads to the operation on a cube and 3 censi equal to 5, and the second on a cube and 6 censi and 8 cose equal to 1000. [That is, $x^3 + 3x^2 = 5$; $x^3 + 6x^2 + 8x = 1000$.] By F. Luca and others these equations have up to now been considered to be impossible of solution by a general rule. You believe that with such questions you can place yourself above me, making it appear that you are a great mathematician. I have heard that you do this towards all the professors of this science in Brescia, so that they for fear of these your questions do not dare to talk with you; and perhaps they know more about this science than you.

M. Z.— I understand as much as you have written to me and that you think such cases are impossible.

N.— I do not say such cases are impossible. In fact, for the first case, that of the cube and the censi equal to a number, I am convinced I have found the general rule, but for the present I want to keep it secret for several reasons. For the second, however, that of the cube and censi and cose equal to a number, I confess I have not up till now been able to find a general rule; but with that I do not want to say it is impossible to find one simply because it has not been found to the present. However, I am willing to wager you 10 ducats against 5 that you are not able to solve with a general rule the two questions that you have proposed to me. And that is something for which you should blush, to propose to others what you yourself do not understand, and to pretend to understand in order to have the reputation of being something great.”

That ends the first encounter.

5

We now go back a ways to the aforementioned pupil of Scipio del Ferro, Antonio Maria Fior, sometime of Venice. He seems to have turned Ferro’s formula to account in the popular mathematical contests then in vogue. Hearing of Tartaglia’s claim to solving some cubic, possibly publicized through Da Coi, and thinking Tartaglia an impostor and himself knowing Ferro’s solution of $x^3 + px = q$, he challenged the latter to a contest. It was set for February 22, 1535. Tartaglia, knowing Fior was only an arithmetician, gave the contest little thought at first. But when he heard that “a great master” “30 years ago” had communicated to him the solution of a cubic, he became worried and set himself to study the equation $x^3 + px = q$. (He already had solved $x^3 + px^2 = q$.) On February 14, eight days before the date set for delivering the solutions to the notary who kept the stakes, he found the solution of $x^3 + px = q$; and on the next day he also found the solution of $x^3 = px + q$.

Each had challenged the other with thirty questions. As Tartaglia had suspected, all Fior’s problems reduced to the form $x^3 + px = q$, and he solved them all in two hours. It almost seemed wicked of Tartaglia, for he had constructed problems such that most of them led to the solution of $x^3 + px^2 = q$ and Fior could not answer a one of them. “I waived the stake and took the honor,” says Tartaglia.

Thus ended the second encounter.

We read of these things in the histories. But our modes of life and thinking, our physical environment, are so removed from 16th century Italy that it is hard for us to reconstruct the tenseness and excitement that accompanied these contests. Honor, gold, and the instinct of game were powerfully present. The questions themselves—the instruments of combat—what did they look like? The histories tell us about $x^3 + px = q$. That seems so general and colorless. And then there were no $x^3 + px = q$. There were “cube and cose equal to a number” and similar expressions. The challenges did not come in that form either—they came as problems. And since this is a sidelight, we shall see what they are. And, gentle reader, so as to be along in spirit with the tense partisans of that February 22, 1535, solve one or two; you are along in the opening skirmish of the famous “Battle of the Cubic” of the 16th century.

These were the questions submitted by Fior for February 22, 1535:

1. Find the number which added to its cube root gives 6.
2. Find two numbers in double proportion $[x, 2x]$ such that if the square of the larger is multiplied by the lesser and to the product is added the sum of the numbers, the result is 40.
3. Find a number which added to its cube gives 5.
4. Find three numbers in triple proportion $[x, 3x, 9x]$ such that if the square of the smallest is multiplied by the largest and the product be added to the mean number, the result is 7.
5. Two men were in partnership, and between them they invested a capital of 900 ducats, the capital of the first being the cube root of the capital of the second. What is the part of each?
6. Two men together gain 100 ducats. The gain of the first is the cube root of the gain of the second. What is the gain of each?
7. Find a number which added to twice its cube root gives 13.
8. Find a number which added to three times its cube root gives 15.
9. Find a number which added to four times its cube root gives 17.
10. Divide fourteen into two parts such that one is the cube root of the other.

11. Divide twenty into two parts such that one is the cube root of the other.
12. A jeweler buys a diamond and a ruby for 2000 ducats. The price of the ruby is the cube root of the price of the diamond. Required the value of the ruby.
13. A Jew furnishes capital on the condition that at the end of the year he shall have as interest the cube root of the capital. At the end of the year the Jew receives 800 ducats, as capital and interest. What is the capital?
14. Divide thirteen into two parts such that the product of these parts shall equal the square of the smallest part multiplied by the same.
15. A person buys a sapphire for 500 ducats and gains the cube root of the capital invested. What was his gain?

16–30 deal with lines, triangles, and various equilateral polygons with sides so divided as to become problems of dividing 7, 12, 9, 25, 26, 28, 27, 29, 34, 12, 100, 140, 300, 810, 700 each into two parts such that one is the cube root of the other.

As we see, all these reduce to the form $x^3 + px = q$.

Of Tartaglia's 30 challenge questions to Fior we have record of only the first four. These follow:

1. Find an irrational quantity such that when it is multiplied by its square root augmented by 4, the result is a given rational number.
2. Find an irrational quantity such that when it is multiplied by its square root diminished by 30, the result is a given rational number.
3. Find an irrational quantity such that when to it is added four times its cube root, the result is thirteen.
4. Find an irrational quantity such that when from it one subtracts its cube root, the result is 10.

These problems resolve themselves into solving for the irrational quantity x in

$$\begin{aligned} x^3 + mx^2 &= n; & m^2x^2 &= x^3 + n; \\ x^3 + mx &= n; & x^3 &= mx + n. \end{aligned}$$

6

It is the ever-moving Da Coi again who brings in the next important personage in these events, Giro-

lamo Cardano. For after his interview with Tartaglia he leaves Brescia and moves to Milano. There he meets Cardan who engages him to instruct one of his classes. Da Coi tells him about Tartaglia and his discovery. Cardan, at this time preparing material for his ambitious work, *Ars Magna*, was much interested in the mathematical duel of Tartaglia and Fior. He therefore sends as messenger Zuan Antonio de Bassano, a book seller, to Tartaglia to inquire about his invention. The atmosphere of the time and the temperament of the principals are well sketched by Tartaglia, under the date of January 2, 1539, in *Quesito XXXI. Fatto da M. Zuanantonio libraro, per nome d' un Messer Hieronimo Cardano, Medico et delle Mathematiche lettor publico in Milano, adi. 2. Genaro, 1539.*

Zuanantonio—Messer Nicolo, I have been directed to you by a certain man, a good physician in Milano called Messer Hieronimo Cardano, who is a great mathematician and gives public lectures on Euclid in Milano; at present he has a work in press on the art of arithmetic, geometry, and algebra, which will be a beautiful thing. He has heard of the contest you had with Maestro Antoniomaria Fiore and how you agreed to prepare 30 cases or questions each, and that you did that. And his Excellency has heard that all the 30 questions proposed to you by Maestro Antoniomaria led you by algebra to an equation of the cosa and the cube equal to a number (*che ui conduceano in Algebra in un capitolo di cosa e cubo equal a numero*), and that you found a general rule for such an equation and that by the power of this invention you solved in the space of two hours all the 30 cases he proposed to you. On this account his Excellency begs that you would be so kind as to send him this rule that you have invented; and if it pleases you he will insert it in his forthcoming book under your name.

N.—Tell his Excellency that he must pardon me; that when my invention is to be published it will be in my own work. His Excellency must excuse me.

Z.—If you do not want to impart your invention to him, his Excellency ordered me to ask you at least to let him have the 30 abovementioned cases which were proposed to you together with their solutions [meaning the results, not the rule obtaining them].

N.—Not even that can be. For whenever his Excellency observes one of these cases and its solution he will get to understand the rule that I found. And by means of this one rule many others dealing with this subject can be derived."

So far, Tartaglia.

After this second rebuff Zuanantonio proposes seven problems leading to these equations:

$$2x^3 + 2x^2 + 2x + 2 = 10 \quad (1)$$

$$2x^3 + 2x^2 + 2x + 2 = 10x \quad (2)$$

$$2x^3 + 2x = 10 \quad (3)$$

$$2x^3 + 2x^2 = 10 \quad (4)$$

$$2x^3 + 2 = 10x \quad (5)$$

$$x^4 + 8x^2 + 8^2 = 10x^3 \quad (6)$$

$$x^3 + 3x^2 = 21 \quad (7)$$

Somewhat hotly Tartaglia rejoins: “These questions are from Messer Zuanne da Coi. And from no one else, for I recognize the last two. Two years ago he proposed to me a question like the sixth and I made him own up that he neither understood the problem nor knew the solution. He also proposed one similar to the last one, which involves working in census and cubes equal to a number [that is, $x^2 + px^3 = q$] and out of my kindness I gave him the solution less than a year ago. For such solutions I found a particular rule applicable to similar problems.”

The bookseller maintains the questions are Cardan’s, however. And to support his request he praises Cardan’s abilities and deftly mentions his connection with a certain high and rich personage, the Marquis del Vasto, a benefactor who was to publish Cardan’s book.

“I do not say his Excellency is not a very learned and capable person”, says Tartaglia. “But I say he is not able to solve the seven problems which have been proposed to me to be solved by a general rule.” When the messenger leaves he gives him a copy of Fior’s 30 questions but not the solutions.

Cardan’s reply, February 12, 1539, is full of bitter insult. You are not at the top of the mountain, you are only at its foot, in the valley, he tells Tartaglia, in substance. It is peculiar that you attribute the seven problems to Da Coi, as if there were no one in Milano able to do such a thing. Da Coi is as young as he says he is; I have known him since before he could count to ten. You said if one of these problems is solved, they all are solved. That is completely wrong. I wager 100 ducats you are not able to reduce them to one, nor to two, nor to three equations. (This is the purest invention of Cardan: Tartaglia had said nothing of the kind. Or else the bookseller had misunderstood him.) Concluding, he proposes two problems. The first, taken from Pacioli but not solved by him, requires to find four numbers in geometric pro-

gression whose sum is 10 and whose square sum is 60. The second concerns two men in partnership who gain the cube of the tenth part of their several capitals.

To the first Tartaglia in his restrained reply of February 18 gives an elegant solution. But he is not cajoled into giving away his secret by solving the second, still keeping to himself his solution of “the cube and the cose equal to a number”.

Neither tricks nor insults succeeding, Cardan turns to flattery and praise. So in a letter dated March 13, 1539 he begins: “Messer Nicolo, mio carissimo.” Asks Tartaglia not take his former words up in bad part. Blames it onto Da Coi who had given him a wrong idea of Tartaglia. Now the ungrateful wretch has left Milano unceremoniously and also left the sixty pupils he had secured for him. He ends by inviting Tartaglia to visit him in Milano and says that the Marquis del Vasto is anxious to meet him. (This was probably pure fiction.) He concludes the letter with high praise for the nobleman and warm feelings for “mio carissimo” Tartaglia:

“And so I urge you to come at once, and do not deliberate; for the said Marquis is a remunerator of all virtuosi, so liberal and magnanimous that no one who serves his Excellency in any matter remains unsatisfied. So do not hesitate to come, and come and live in my house and no otherwheres. May Christ keep you from harm. March 13, 1539. Hieronimo Cardano, Physician.”

This was the rift in the wall that made Tartaglia’s citadel crumble. He accepts the invitation and stays a few days in Cardan’s house. Their conversation is recorded in Quesito XXXIV under date March 25, 1539.

C.— It is convenient for us that the Marquis has just left for Vigevano so we can talk about our affairs till he returns. You surely have not been any too obliging in not showing me your solutions of the cube and cose equal to a number that I have so earnestly asked you to do.

T.— I tell you, I am niggardly in this matter, not for the sake of this simple equation only and the things that it has enabled me to find, but for the sake of all the things this equation ought to help me cover in the future. For it is a key that opens up the investigation of a great many other equations. If I were not now occupied with the translation of Euclid (I am already on Book XIII) I would already have discovered a general rule for many other equations. [Then he discusses his plan for his book on algebra.] If I now showed the solution to a speculative mind,

like your Excellency, he could easily discover the other solutions and publish them as his own, which will completely spoil my project. [Notice all along the distinction between solutions, answers, and the formula or process that gives the solutions.] This is the reason that has compelled me to be so discourteous toward your Excellency; so much the more since you are about to publish a work on a similar subject and in which work, you wrote you would like to insert my invention under my name.

C.— But I wrote you that if that did not meet your approval, I will promise to keep it a secret.

T.— As to that, I just can't believe you.

C.— Then I swear you by the holy Evangelists of God and as a man of honor that I will not only never publish it, but I will write for myself in code so that no one finding them after my death understand. If you will now believe me, believe; if not, let it pass.

T.— If I did not believe such an oath, I should certainly be regarded as a man without faith. But I have decided to go to Vigevano to find the Marquis; for I have already been here three days and am tired of waiting. On my return I promise to reveal it all.

C.— If you wish to see the Marquis I will give you a letter so that he may know who you are. But before you go I wish you would show me the rule, as you promised.

Then Tartaglia gives him the solution for $x^3 + px = q$ and $x^3 + q = px$. Instead of a code Tartaglia gives it in twenty-five lines of poetry, seven tercets followed by a quatrain. It must have been as good as a code, for in a letter of April 9th Cardan has trouble with this mathematical poetry. In his reply Tartaglia says it is not $ut = \frac{1}{3}p^3$, but $ut = (\frac{1}{3}p)^3$.

We will just give a taste of this mathematical poetry by quoting one tercet:

Quando che'l cubo con le cose appresso,
Se aggruaglia a qualche numero discreto
Trouan dui altri, differenti in esso.

Meaning: If $x^3 + px = q$, let $t - u = q$.

The next few lines say: Also let $ut = (\frac{1}{3}p)^3$, then $x = \sqrt[3]{t} - \sqrt[3]{u}$.

Tartaglia must already have begun to feel uneasy, for on leaving Cardan he says to himself: "I will not go to Vigevano. I will go back to Venice, come what may." In the exchange of letters that follow it becomes evident that Cardan is putting his powerful mind to work on Tartaglia's formulae from every angle and soon discerned implications that Tartaglia himself had either not been able to see or was too

busy to follow up (he was busy with his translation of Euclid).

The Irreducible Case came up in a letter from Cardan in August 1539, when Cardan asks: How about $x^3 = px + q$ when

$$\left(\frac{p}{3}\right)^3 > \left(\frac{q}{2}\right)^2$$

as in $x^3 = 9x + 10$?

Tartaglia saw that Cardan was now making his own investigations and felt none too good about it. He himself could not solve the difficulty, and his answer to Tartaglia lacks frankness. "Has Tartaglia lost spirit maybe from much studying and reading?" banters Cardan in his next letter. "If he is sure of understanding the rule he will wager 100 ecus against 25 that he can solve $x^3 = 12x + 20$." Tartaglia did not answer.

On January 5, 1540 came a noteworthy letter from Cardan— noteworthy in the light of what followed. Very friendly; not "mio carissimo" now, but "quanto fratello". "That devil" Zuanne dal Colle (as Cardan always spelled it) has returned to Milano and caused him no end of grief. Both in his teaching and in other matters. But, warns Cardan, he has your equation $x^3 + px = q$ and $px + q = x^3$, and he boasts that during his sojourn in Venice he had a discussion with Fior and so arrived at what he searched for. Then he tells of various algebraic solutions that Zuanne had solved, giving details. The whole letter looks like a build-up for 1545, to show that the knowledge of the cubic was not Tartaglia's only.

But Tartaglia does not catch the drift. "Cardan has a mind more dull than I thought", he muses. "Zuanne imposes on him when he says that he has the solution of equations. But I do not want to reply. I have no more affection for him than I have for Messer Zuanne. I will leave them to one another. But I can see that he has lost spirit and does not see how things will turn out."

Then all correspondence between these two ceases.

The next five years were quiet, seemingly. Tartaglia, busy with his translations of Euclid and Archimedes, holding in abeyance his future work on algebra; Cardan, assisted by the brilliant Ferrari, working on the *Ars Magna*. In 1545 this monumental work appeared from the Nürnberg press of Petreius. In it, with his consent and under his name, was Ferrari's solution of the biquadratic. In it, and with his name, but not with his consent, was Tartaglia's solution of $x^3 + px = q$. The solution that was to have

been written in code lest the world should get knowledge of it was broadcast on the pages of Cardan's most ambitious work.

It is given as Chapter XI of the *Ars Magna*, and is prefaced by a statement that Scipio Ferro had first found the solution, that later Tartaglia also invented it and showed the solution, but not the demonstration, to Cardan. Tartaglia had told Cardan he was jealous of the solution of $x^3 + px = q$ not so much for the equation itself, but for the work to which it was the key. And true enough, using this key, ten additional chapters on the cubic besides Ferrari's on the biquadratic, enrich the contents of the *Ars Magna*. How Tartaglia felt when his eyes saw this, we can imagine. Or can we?

Tartaglia's reply to the statement and the act is given in his *Quesiti*—documented with names, circumstances, dates, places,—published the following year. Cardan never satisfactorily met those accusations in writing, nor could Tartaglia entice him to meet him in person for a mathematical combat.

7

Now comes a most unique spectacle in mathematical history; not as mathematics but as human passions, wickedness, and contrariness.

Cardan did not reply to the accusations of Tartaglia. But Ludovico Ferrari, his grateful pupil, "suo creato", took up the gauntlet for his master. On February 10, 1547, he sent a public challenge to Tartaglia at Venice: a pamphlet with four pages of content and four (!) pages of names of mathematicians in various universities and cities to whom copies of the challenge had been sent, fifty in all. Among these we notice the name of Ferro's successor, "Hannibal dalle Nave"; but neither Da Coi nor Fior.

"Messer Nicolo Tartaglia", it begins, "there has come into my hands a book by you called *Quesiti ed inventioni nuovi*, in the last treatise of which you mention his Excellency Signor Hieronimo Cardano, a physician at Milano, who is at present a public lecturer in medicine at Pavia. And you are not ashamed to say that he is ignorant in mathematics, that he is a dull individual, deserving to have Messer Giovan da Coi placed before him ... I think you have done this, knowing that Signor Hieronimo has such a great genius that not only in medicine, which is his profession, has he a reputation for ability, but also in mathematics, in which he has at times indulged as a game, to get recreation and enjoyment,

and in which he has become so widely successful that without exaggeration he is considered one of the great mathematicians."

Besides a multitude of errors, the challenge continues, Tartaglia has also plagiarized from Jordanus, whose propositions he has placed in the 8th book without citing his name. Tartaglia has blamed Cardan unjustly; he, who is not worthy to mention Cardan's name (*il quale à pena sete degno di nominare*). Thereupon he challenges Tartaglia to a public disputation from ancient and modern authors on "Geometry, Arithmetic, and all the disciplines that depend on these, as Astronomy, Music, Cosmography, Perspective, Architecture, and others, ... and not only on what Latin, Greek and "vulgar" [vernacular, modern] authors write on these subjects, but also on your own inventions."

The time was thirty days; the stake, up to 200 *scudi*, to be decided by Tartaglia. "And in order that this invitation shall not appear too private, I have sent a copy of this writing to each of the gentlemen named below."

Thus begins the fourth part of this celebrated controversy.

Tartaglia replies on February 19, nine days later. Also a printed pamphlet, equally formal: "From Nicolo Tartaglia of Brescia, Professor of mathematics in Venice, to Messer Ludovico Ferraro, Public Lecturer of Mathematics in Milano." Six pages of compact print, with four witness signatures. But instead of an impressive list of mathematicians at the end, he has a postscript:

"And in order that this reply of mine shall not appear too private, I have had 1000 copies printed to send them around Italy in general; since I am not acquainted with the cities in Italy or with the universities, where one can buy the friendship and knowledge of experts and scholars, as you do (because, in truth, my experience and acquaintance are limited to my study and to my students). For this reason I do not only not have the friendship but not even the acquaintance of these persons."

Therefore, he continues, he will not send his reply to the persons Ferrari names; could not do it even to a certain named person, "for he died two months ago." (Ferrari's challenge had come nine days ago!)

As for a response, he will not meet Ferrari in combat. It is with Cardan he has a quarrel, and when that gentleman is ready, Tartaglia will accept. So he sends a counter-challenge that Cardan and Ferrari on one side and he on the other submit one to the other a list of problems to solve.

Six weeks later, on April 1, comes a second challenge from Ferrari—and *this time in Latin*. Why change from Italian to Latin? Tartaglia thought he knew.

In Cartello II Ferrari touches upon the solution of the cubic equation. Tartaglia has taken umbrage at Cardan for publishing his solution of the cubic. What if the published solution was that of a third party? Five years ago, declares Ferrari, in 1542, Cardan and Ferrari were in Bologna and there visited Annibale della Nave, Scipio Ferro's son-in-law, who showed them books written by Scipio; and there was the solution Cardan published. Annibale is alive today and can be called as a witness anytime. (This would be a serious argument, except it is not convincing. Why, since there was no secrecy about it, had Cardan in *Ars Magna* not mentioned Scipio's writings and Annibale's part, instead of referring only to Scipio and Fior?)

A large portion of Tartaglia's replies is sparring for objectives. He regularly wants to get into combat with Cardan himself; just as regularly the slippery Ferrari turns him off. Another objective is to have the contest be a list of challenge questions, to be solved in a specific time; Ferrari wants a public disputation in Rome, Florence, Pisa, or Bologna, to be chosen by Tartaglia, and judges to be selected from persons in the city chosen.

Two points taken up in Risposta II evoke our sympathy:

The first concerns the mode of contest. Besides considering the method of challenge lists a better arbiter of ability than public disputations, the sender may have had an added reason for his choice. Tartaglia, "the stammerer", had an impediment of speech ever since as a child he was cut by a French soldier in the cathedral massacre at Brescia. In a public disputation he would have a serious handicap engaging the oily Cardan and the brilliant Ferrari.

The second dealt with where and with whom the stake was to be deposited *and if only gold was to be used*, or whether Tartaglia could, for part of the sum, deposit printed copies of the *Quesiti*. Remembering Tartaglia's worldly circumstances one can here read much between the lines.

Replying to the charge of plagiarism, he says: Though the statement of the propositions emanated from Jordanus, the demonstrations and arrangement were Tartaglia's. The statement of a proposition without the proof is of no value. Answering Ferrari's reference to the cubic: "It would be presumptuous of me to give the impression that the result which

I discovered could not also have been discovered at other times and by other persons, and that they cannot likewise be discovered in the future and by other persons: even when they will not be given to the public by Signor Hieronimo or myself. But this can I say with truth, that I never saw these things in any author but discovered it myself."

He pokes a thrust at Cardan for not being willing to enter the contest but sending Ferrari instead. And then he dishes up this pretty one:

"You say that you have heard that in the past few years I have made machines and invented several types of instruments and that people think that by my persistent knowledge I have succeeded in making a machine with which I can shoot clear to Milano while I am stationed in Venice.

"Regarding this particular, I answer they are not at all wrong. For since the presentation of your Cartello I have actually built one with which, while I am in Venice, I can shoot, not only as far as Milano, but even as far as Pavia [Cardan had left Milano and was now located in Pavia], and shoot with such a direct aim that I will not only scare you and Signor Hieronimo but cause you great anguish."

Then he proposes a challenge of thirty-one questions. He states that he can solve them, adding: "I am not like Signor Hieronimo who presents cases he does not know how to solve himself."

Ferrari counters with 31 other questions in the Cartello III (June 1) without answering Tartaglia's. This letter is a highly insulting piece. "In response to my reply," he says, "I have received your *tartagliata* [a pun on the etymology of Tartaglia's name]: which, though long and confused, contains nothing but insults, refusals to admit defeat, and a fixed idea of wanting to fight while running away."

In Risposta III (July 9) Tartaglia gives the solution to Ferrari's 31 questions and boasts he is the victor. In Cartello V (October, 1547) Ferrari tears Tartaglia's answers to pieces mercilessly and claims only five are correct. This Cartello is almost a book in size, all of 55 pages; 41 pages are mathematics and contain Ferrari's solutions of Tartaglia's challenge list.

Eight months pass before the answer comes, the longest time between any of these exchanges. And then the unlooked-for happens: on June 1, 1548, Tartaglia accepts the challenge to a public disputation and even to have it in Cardan's and Ferrari's own bailiwick, Milano. How and why the change? He may have become so exasperated with Ferrari's manhandling of his solutions in Cartello V and of

other provocative matters in the Cartello—and he did want a duel with Cardan—that he was willing to forego his preferred mode of question lists. Or there may have been other reasons as insinuated by Ferrari, necessitating “Brescia *versus* Milano”. For city pride and championships did not begin in the 20th century. Tartaglia had recently moved to Brescia where some of its chief citizens had promised him liberal remuneration for giving public lectures on Euclid. Ferrari insinuates that his acceptance of the challenge to Milano was one of the stipulations.

Tartaglia accepted the challenge. But let no one think there was sportsman’s etiquette there. Cartello VI (July 14, 1548) and Rispona VI (July 24, 1548) perforce takes up arrangements on the business element. But they have plenty of room for smarting sentences, especially Ferrari’s. He seems to have fed on his own anger and vindictiveness.

For a year and a half these tirades had continued. On August 10, 1548, the “disputation” took place, and with what outcome we read in Tartaglia’s own account, written nine years later, in an article interpolated in his *General Trattato*.

It follows:

“In 1547 Cardan and his creature Ludovico Ferraro brought me a challenge in two printed pamphlets. I addressed to them 31 questions, with the stipulation that they should be solved in 15 days after receiving them. After that the solution should be considered as not arrived. They let two months pass without giving any sign of existence, and then they sent me 31 questions without giving me the solutions of any of mine; besides, the terms stipulated had passed by more than 45 days. I found the solution of 10 of them the same day, the next day a few more, later all the rest, and, so as not to let pass the interval of 15 days, I hurried to get them printed and sent to Milano. In order to conceal their slowness in answering my questions or at least a few of them they took up my time with other matters full of silly foolishness, and it was not till the end of seven months that they sent me a public reply where they boasted that they had solved my questions. However, even had that been true, those solutions given so long a time after the term fixed would have been without any merit; but the greater part of them were completely wrong. I desired to proclaim publicly that they were wrong, so, being in Brescia and hence in the neighborhood of Milano, I sent to them both a printed challenge asking them to meet me the following Friday, August 10, 1548 at 10 o’clock near the church called the Garden of the

Order of Zoccolante to argue publicly my refutation of their pretended solutions. Cardan, so as not to be at the examination, left Milano hurriedly.

“On the day set Ferraro came alone to the meeting-place, accompanied by a crowd of friends and by many others. I was alone with my brother whom I had taken along from Brescia. I went before the entire crowd and began by giving briefly an exposition of the subject for discussion and the reason for my arrival in Milano. When I wanted to take up the refutations of the solutions I was interrupted for a period of two hours by words and actions under pretext that there should be chosen, at that very place, a certain number of judges from the auditors present, all friends of Ferraro and to me entirely unknown. I would not consent to such a trick and said that my understanding was that all the auditors were judges, the same as those who read my refutation when printed. Finally they let me speak, and in order not to tire my audience I did not commence with tedious topics from number theory and geometry, but it seemed to me suitable to refute their solution of a question in Chapter 24 in Ptolemy’s *Geography*; and I constrained Ferraro publicly to own that he was in error. When I wished to continue they all began to shout that now I should discuss my own solutions of the 31 questions that had been proposed to me and which I had solved in 3 days. I objected that they should let me finish what concerned my refutations, then I would take up what they asked for. Neither reasoning nor complaining could be heard. They would not let me speak further and gave the word to Ferraro. He began by saying I had not been able to solve the fourth question in Vitruvius, and he expatiated on this clear till the supper hour. Then everybody left the church and went home.”

So far Tartaglia. Seeing the temper of the crowd and fearing violence, he did not wait to continue his disputation the next day but hurriedly left for Brescia by another road than that by which he had come, glad to keep his life.

So ended this combat at the Church of Zoccolante, original even for this age.

8

In 1556 (ten years after the appearance of *Quesiti*), came the first two parts of Tartaglia’s great life work, for the contents of which he had reserved many discoveries made even before the *Inventioni* of 1546.

It was the *General Trattato di numeri, et misure*, a huge, ambitious and well-written work on arith-

metic and algebra. It is said to be the best work on arithmetic in the entire century. The third part, which was not published till 1560, was left uncompleted; for Tartaglia died in 1557. It was largely algebra and it is thought the last division was to have included his work on the cubic equation. As it is, we have only so much of his work in this field as is found in the *Quesiti* of 1546.

Reading Bombelli's x -purgated Algebra

ABRAHAM ARCAVI and MAXIM BRUCKHEIMER

College Mathematics Journal 22 (1991), 212–219

Reading mathematics is hard work and reading a four hundred year old mathematics text is four hundred times harder. The language, notation and also the spirit are different from ours. If the reader is not already convinced from past experience, the following extract should prove the point.

Let us first assume that if we wish to find the approximate root of 13 that this will be 3 with 4 left over. This remainder should be divided by 6 (double the 3 given above) which gives $\frac{2}{3}$. This is the first fraction which is to be added to the 3, making $3\frac{2}{3}$ which is the approximate root of 13. Since the square of this number is $13\frac{4}{9}$, it is $\frac{4}{9}$ too large, and if one wishes a closer approximation, the 6 which is the double of the 3 should be added to the fraction $\frac{2}{3}$, giving $6\frac{2}{3}$, and this number should be divided into the 4 which is the difference between 13 and 9, . . .

If you understood—don't read on. If you didn't, then in this article we illustrate in some detail how a little perseverance can turn "obscurity" into a rewarding experience for students.

1 Bombelli's method

The text quoted above was written by Rafael Bombelli, a 16th century Italian mathematician. He wrote an important textbook which appeared in two editions, *L'algebra parte maggiore dell'aritmetica* (1572) and *L'algebra* (1579). It includes Bombelli's contributions to the solution of cubic and quartic equations and many geometrical problems solved algebraically (see, for example, [2] and [4]).

In this paper we reproduce from the 1579 edition the algorithm he introduced to approximate square roots, which is the subject of the obscure paragraph

quoted in our introduction. At the time Bombelli wrote his *Algebra*, decimal fractions were not yet in use; they were introduced by Simon Stevin in his book *La disme* ("The tenth") in 1585. Bombelli developed his method using common fractions. We will follow in his footsteps. The chapter starts (the English version from which we quote is [5, p. 81]) with a very human touch—not exactly in the style

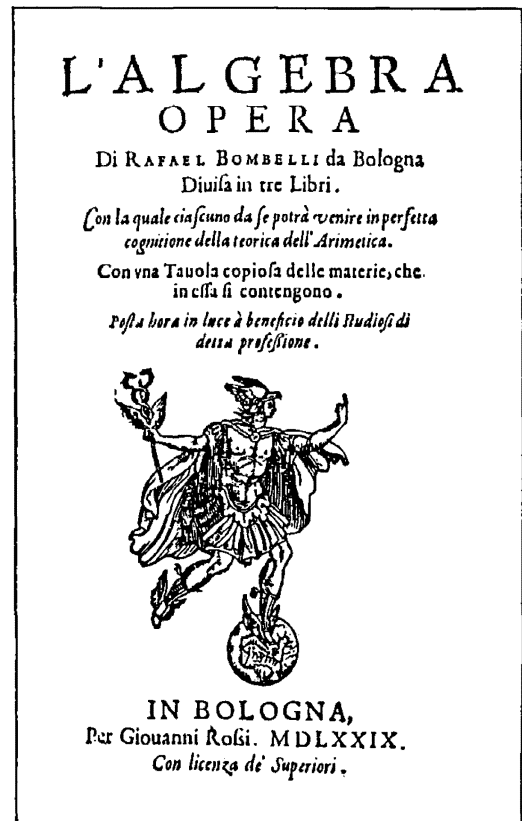


Figure 1.

of a modern textbook – serving precisely the purpose of engaging and motivating readers.

Method of Forming Fractions in the Extraction of Roots (*Modo di formare il rotto nella estrazione delle Radici quadrate*)

Many methods of forming fractions have been given in the works of other authors; the one attacking and accusing another without due cause (in my opinion) for they are all looking to the same end. It is indeed true that one method may be briefer than another, but it is enough that all are at hand and the one that is the most easy will without doubt be accepted by men and be put in use without casting aspersions on another method. . . . In short, I shall set forth the method which is the most pleasing to me today and it will rest in men's judgement to appraise what they see: meanwhile I shall continue my discourse going now to the discussion itself.

The above extract is followed by the text we quoted in our introduction. Even if we translate the “recipe” for extracting square roots described there from its rhetorical form (see, for example, [6, pp. 378–379]) into modern symbols, and follow the calculations, the procedure is still unmotivated. Why should one divide the remainder by 6? Why should one add the 6 to the $\frac{2}{3}$? Why should one divide it into 4? And so on.

Why did Bombelli find this method *the most pleasing*? It would seem to be hard to agree with him, at least when one reads it for the first time. However, in a subsequent paragraph, Bombelli himself provides a fuller explanation of his method, which we reproduce (the English version we quote here is taken from [3, pp. 69–70].) followed by the modern notation.

1. *Let us suppose we are required to find the root of 13. The nearest square is 9, which has root 3. I let the approximate root of 13 be 3 plus 1 tanto [unknown].*

$$3 + x = \sqrt{13}$$

2. *Its square is 9 plus 6 tanti p. 1 power. We set this equal to 13.*

$$(3 + x)^2 = 9 + 6x + x^2 = 13$$

3. *Subtracting 9 from either side of the equation we are left with 4 equal to 6 tanti plus one power.*

$$6x + x^2 = 4$$

4. *Many people have neglected the power and merely set 6 tanti, equal to 4. The tanto then comes*

Modo di formare il rotto nella estrazione delle Radici quadrate.

Molti modi sono stati scritti da gli altri autori de l'uso di formare il rotto; l'uno cassando, e accusando l'altro (al mio giudicio) senza alcun proposito, perche tutti mirano ad un fine; E ben vero che l'una è più breue dell'altra, ma basta che tutte suppliscono, e quella ch'è più facile, non è dubbio ch'essa farà accettata da gli huomini, e sarà posta in uso senza cassare alcuno; perche potrà essere, che hoggi io insegnassi una regola, laquale piacerebbe più dell'altre date per il passato, e poi venisse un'altro, e ne trouasse una più vaga, e facile, e così sarebbe all'hora quella accettata, e la mia confutata, perche (come si dice) la esperienza ci è maestra, e l'opra loda l'artefice. Però metterò quella che più à me piace per hora, e farà in arbitrio de gli huomini pigliare qual vorranno: dunque venendo al fatto dico. Che presuposto, che si voglia il prossimo lato di 13, che farà 3, e auanzerà 4, il quale si partirà per 6 (doppio del 3 sudetto) ne uiene $\frac{2}{3}$, e questo è il primo rotto, che si hà da giungere al 3, che fa $3\frac{2}{3}$, ch'è il prossimo lato di 13, perche il suo quadrato è $13\frac{4}{9}$, ch'è superfluo $\frac{4}{9}$, ma uolendosi più approssimare, al 6. doppio del 3 se gli aggiunga il rotto, cioè li $\frac{2}{3}$, e farà $6\frac{2}{3}$, e per esso partendosi il 4, che auanza dal 9 fino al 13,

Figure 2.

to $\frac{2}{3} \dots$

$$6x = 4 \implies x = \frac{2}{3}$$

5. . . . and the approximate value of the root is $3\frac{2}{3}$ since it has been set equal to 3 p. 1 tanto.

$$\sqrt{13} \approx 3 + x \approx 3\frac{2}{3}.$$

So far, Bombelli has found a first approximation by “neglecting” (*lasciato andare*) the value of x^2 and thus obtaining $x = \frac{2}{3}$, the fraction that is to be added to 3 to obtain an approximation to $\sqrt{13}$. The second approximation is now found by taking into account what was neglected in the first approximation.

6. However, taking the power into account, if the tanto is equal to $\frac{2}{3}$, the power will be $\frac{2}{3}$ of a tanto, which, added to the 6 tanti, will give us 6 and $\frac{2}{3}$ tanti, which are equal to 4.

$$\left. \begin{array}{l} 6x + x^2 = 4 \\ 6x + xx = 4 \end{array} \right\} \implies 6x + \frac{2}{3}x = 4$$

7. So the tanto will be equal to $\frac{3}{5}$, and since the approximate is 3 p. 1 tanto it comes to $3\frac{3}{5}$.

$$\implies x = \frac{4}{6 + \frac{2}{3}} = \frac{3}{5} \implies 3 + x = 3\frac{3}{5}.$$

Note that Bombelli has taken the product $x^2 = xx$, and given one x the value previously obtained, and the other becomes the new fraction that is to be added to 3 to obtain the next approximation to $\sqrt{13}$. He now uses this *double-entendre* for x recursively.

8. But if the tanto is equal to $3\frac{3}{5}$, the power will be $\frac{3}{5}$ of a tanto and we obtain $6\frac{3}{5}$ tanti equal to 4 . . .

$$\left. \begin{array}{l} 6x + x^2 = 4 \\ 6x + xx = 4 \end{array} \right\} \implies 6x + \frac{3}{5}x = 4 \dots$$

Solving for x , Bombelli obtains a third approximation, $3\frac{20}{33}$.

The process can be continued for as long as one has patience. Or to put it in Bombelli's words: *e cosi procedendo si puo approssimare a una cosa insensibile* (and this process may be carried to within an imperceptible difference [5]).

2 Bombelli's method and continued fractions

Bombelli's algorithm can also be described in the following way. We wish to find an x so that $\sqrt{13} = 3 + x$. Squaring both sides, $13 = (3 + x)^2$. Whence $6x + x^2$ or $x(6 + x) = 4$, and finally $x = 4/(6 + x)$.

Note that in the expression $4/(6 + x)$, the x takes the value obtained in the previous stage. However, the x on the left takes a new value obtained in the present stage. This corresponds to the *double-entendre* noted above. In more precise modern notation we would write

$$x_{n+1} = \frac{4}{6 + x_n}.$$

Using our notation we obtain the continued fractions

$$\begin{aligned} x_2 &= \frac{4}{6 + x_1}, \\ x_3 &= \frac{4}{6 + x_2} = \frac{4}{6 + \frac{4}{6+x_1}}, \\ x_4 &= \frac{4}{6 + x_3} = \frac{4}{6 + \frac{4}{6 + \frac{4}{6+x_1}}}. \end{aligned}$$

One cannot assert that Bombelli invented continued fractions, since this form—or anything that could suggest it—does not appear in his text. However, we read (in [6, pp. 419–420]) that:

Although the Greek use of continued fractions in the case of greatest common measure was

well known in the Middle Ages, the modern theory of the subject may be said to have begun with Bombelli (1572). . . The next writer to consider these fractions, and the first to write them in substantially the modern form was Cataldi (1613), and to him is commonly assigned the invention of the theory. His method was substantially the same as Bombelli's, but he wrote the result of the square root of 18 in the following form:

$$4. \& \frac{2}{8} \& \frac{2}{8} \& \frac{2}{8}$$

3 Bombelli's method in the classroom

In this section we would like to suggest how Bombelli's method can be used in the classroom.

“Dictionary” questions. This activity helps the students to become acquainted with unknown notation, symbols, names of concepts, or formulations in the source. For example, the following dictionary is to be completed by the student, as a first step in deciphering the text.

<i>Rhetorical Language</i>	<i>Modern Notation</i>
p. or plus	—
equal to	—
one quantity (unknown)	—
power (second, of unknown)	—
—	$3 + x$
9 plus 6 tanti p. 1 power	—

The translation process illustrates the immense power which that little x , which we take so for granted, gives us. This is further illustrated in the next section.

“Redoing” and applying the mathematics. Once the terminology is understood, we give students the left side of the eight steps from the section “Bombelli's Method” and ask them to translate each step into modern notation. When the method is clear from the algorithmic point of view, we ask students to apply it to other numbers and thus give the experience some permanence. For example, the calculation for $\sqrt{2}$ gives the continued fraction

$$1 + \frac{1}{2 + \frac{1}{2 + \frac{1}{2 + \dots}}}$$

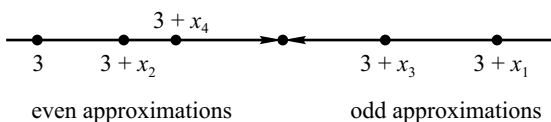
This activity can be integrated into one or more places in the curriculum. The main subject is obviously approximating square roots and Bombelli's method can be compared to others. It can also be discussed when dealing with fractions and continued fractions, as part of a unit on irrational numbers, or as a nice illustration of iteration.

Issues for discussion. We have overcome our perplexity when we first read Bombelli's "recipe". We take the opportunity to point out to students the moral learned from the exercise. Unlike reading in many other areas, reading mathematics, even when the notation is modern, also involves writing, re-doing in other ways, drawing diagrams, and definitely rereading. But, the task is still incomplete. Bombelli's text raises many mathematical issues for the critical reader. We deal with some of them in the classroom in the following way.

Compare the successive approximations of $\sqrt{13}$. Resorting to calculators in order not to place the burden of the activity on comparing common fractions, we find that the first approximations in decimal form are:

$$\begin{aligned} 3 + x_1 &= 3.66666\dots, & 3 + x_2 &= 3.60, \\ 3 + x_3 &= 3.606060\dots, \\ 3 + x_4 &= 3.6055045\dots \end{aligned}$$

Looking at the first four decimal digits of $\sqrt{13} = 3.6055513\dots$ one notices that the even approximations are less than $\sqrt{13}$ and increasing, and the odd approximations are greater than $\sqrt{13}$ and decreasing, as illustrated in the following diagram.



Had Bombelli presented his method today, he would immediately be required to prove that, if one continues the process indefinitely, the sequence will indeed converge to the desired root. A sketch of the proof follows.

The first approximation to $\sqrt{13}$ is greater than $\sqrt{13}$.

$$\left(3 + \frac{2}{3}\right)^2 = 9 + 4 + \frac{4}{9} > 13.$$

The second approximation to $\sqrt{13}$ is less than $\sqrt{13}$.

$$\left(3 + \frac{3}{5}\right)^2 = 9 + \frac{18}{5} + \frac{9}{25} < 13.$$

With mathematical induction, one can prove that every odd approximation is greater than $\sqrt{13}$ and that every even approximation is less than $\sqrt{13}$. The two proofs are similar. We sketch the case for the odd approximations. The induction starts with the step 1 above. Then we assume that $x_{2n-1} > \sqrt{13} - 3$ and prove that x_{2n+1} , the subsequent odd approximation, is also greater than $\sqrt{13} - 3$. In order to do that, we express x_{2n+1} in terms of x_{2n-1} ,

$$\begin{aligned} x_{2n+1} &= \frac{4}{6 + \frac{4}{6+x_{2n-1}}} \\ &> \frac{4}{6 + \frac{4}{6+\sqrt{13}-3}} = -3 + \sqrt{13}. \end{aligned}$$

Next we show that every odd (even) approximation is less than (greater than) its predecessor. For example, in the case of the odd approximations, we have to show that

$$x_{2n-1} > x_{2n+1} \quad \text{or} \quad x_{2n-1} - x_{2n+1} > 0.$$

We again express x_{2n+1} in terms of x_{2n-1} , and, as before, the rest is algebra.

Finally, since the odd (even) approximations are decreasing (increasing) and always greater (less) than $\sqrt{13}$, as shown above, they must approach some limit. We still need to show that the limit l_o of the sequence of odd approximations is the same as the limit l_e of the sequence of even approximations.

One way to show that each limit is $\sqrt{13}$ is the following. Since

$$x_{2n+1} = \frac{4}{6 + \frac{4}{6+x_{2n}}}$$

we can take the limits of both sides to obtain

$$l_e = \frac{4}{6 + \frac{4}{6+l_e}}.$$

Solving the equation we obtain $l_e = \sqrt{13} - 3$. We obtain an identical result for odd approximations, and thus both limits coincide. Bombelli might have provided this argument, making use of his technique of *double-entendre*, had anyone in the 16th century thought that these things needed proving.

4 Final comments

We used this activity with secondary mathematics teachers as a part of a sequence of activities on the historical development of irrational numbers [1]. We found that original sources are a very appropriate

way to convey the feeling of mathematics as a living and developing human endeavor. The rhetorical or quasi-rhetorical expositions, in which the author shows personal preferences (*I shall set forth the method which is the most pleasing to me today*), and presents some arguments in non-rigorous way (*Many people have neglected the power . . .*) were very motivating. And the process of understanding the original source, first at the algorithmic level and then by discussing its mathematical validity as understood with modern eyes, was very enlightening. The teachers found that reading mathematics can be an engaging and enjoyable activity.

References

1. A. Arcavi, M. Bruckheimer and R. Ben-Zvi, History of mathematics for teachers: The case of irrational numbers, *For the Learning of Mathematics* 7 (1987) 18–23.
2. C. B. Boyer, *A History of Mathematics*, Princeton University Press, NJ, 1985, p. 322.
3. P. Dedron and J. Itard, *Mathematics and Mathematicians*, Vol. 2, Transworld, 1974, pp. 69–70.
4. S. A. Jayawardene, The influence of practical arithmetics on the algebra of Rafael Bombelli, *Isis* 64 (1973) 510–523.
5. D. E. Smith, *A Source Book in Mathematics*, McGraw Hill, NY, 1929, pp. 80–82. (The English translation of the Bombelli paragraph is by V. Sanford.)
6. D. E. Smith, *History of Mathematics*, Vol. II, Dover, NY, 1953, 418–420.

The First Work on Mathematics Printed in the New World

DAVID EUGENE SMITH

American Mathematical Monthly 28 (1921), 10–15

1 General description

If the student of the history of education were asked to name the earliest work on mathematics published by an American press, he might, after a little investigation, mention the anonymous arithmetic that was printed in Boston in the year 1729. It is now known that this was the work of Isaac Greenwood who held for some years the chair of mathematics in what was then Harvard College. If he should search the records still further back, he might come upon the American reprint of Hodder's well-known English arithmetic, the first textbook on the subject, so far as known, to appear in our language on this side of the Atlantic. If he should look to the early Puritans in New England for books of a mathematical nature, or to the Dutch settlers in New Amsterdam, he would look in vain; for, so far as known, all the colonists in what is now the United States were content to depend upon European textbooks to supply the needs of the relatively few schools that they maintained in the seventeenth century.

The earliest mathematical work to appear in the New World, however, antedated Hodder and Greenwood by more than a century and a half. It was published long before the Puritans had any idea of migrating to another continent, and fifty years before Henry Hudson discovered the river that bears his name. Of this work, known as the *Sumario Compendioso*, there remain perhaps only four copies, and it is desirable, not alone because of its rarity but because of its importance in the history of education on the American continent, that some record of its contents should be made known to scholars.

In order to understand the *Sumario Compendioso* it is necessary to consider briefly the political and so-

cial situation in Mexico in the middle of the sixteenth century. Cortés entered the ancient city of Tenochtitlan, later known as Mexico, in the year 1519, but its capture and destruction occurred two years later, in 1521. Thus, in the very year that Luther was attacking certain ancient customs and privileges in the Old World, the representatives of other ancient customs and privileges were attacking and destroying a worthy civilization in the newly discovered continent.

The first viceroy of New Spain, which included the present Mexico, was a man of remarkable genius and of prophetic vision — Don Antonio de Mendoza. He assumed his office in 1535, and for fifteen years administered the affairs of the colony with such success as to win for himself the name of “the good viceroy.” He founded schools, established a mint, ameliorated the condition of the natives, and encouraged the development of the arts. In his efforts at improving the condition of the people he was ably assisted by Juan de Zumárraga, the first Bishop of Mexico. Among the various activities of these leaders was the arrangement made with the printing establishment of Juan Cromberger of Seville whereby a branch should be set up in the capital of New Spain.

As a result of this arrangement there was sent over as Cromberger's representative one Juan Pablos, a Lombard printer, and so the “casa de Juan Cromberger” was established, prepared to spread the doctrines of the Church to the salvation of the souls of the unbelievers. Cromberger himself never went to Mexico, but his name appears either on the *portadas* or in the colophons of all the early books. From and after 1545, however, the name is no longer seen, Cromberger having already died in 1540.



Figure 1. Title Page of the First Work in Mathematics printed in the New World

The author of the *Sumario* was one Juan Diez, a native of the Spanish province of Galicia, a companion of Cortés in the conquest of New Spain, and the editor of the works of Juan de Avila, “the apostle of Andalusia,” and of the *Itinerario* of the Spanish fleet to Yucatan in 1518. He is sometimes confused with another Juan Diaz (the name being spelled both ways), a contemporary theologian and author. In a letter written to Charles V in 1533 he is mentioned as a “clérigo anciano y honrado,” so that he must have been advanced in years when the *Sumario* appeared. That this was the case is also apparent from a record of the expedition of 1518 in which it is stated that “triximus vn clerigo que dezia joan diaz,” doubtless a young and adventurous apostle, full of zeal and desire to make known the gospel in the New World.

Juan Diez undertook the work primarily for the

purpose of assisting those who were engaged in the buying of the gold and silver which was already being taken from the mines of Peru and Mexico for the further enriching of the moneyed class and the rulers of Spain. He felt that he could best serve this purpose by preparing such a set of tables as should relieve these merchants as far as possible from any necessity for computation. Apparently, however, he was prompted by the further demand for a brief treatment of arithmetic which should be suited to the needs of apprentices in the counting houses of the New World, and so he devotes eighteen pages to the subject of computation and presents it in a manner not unworthy of the European writers of the period.

The most interesting feature of the work, however, is neither the tables nor the arithmetic; it consists of six pages devoted to algebra, chiefly relating to the

quadratic equation.

The book consists of one hundred and three folios, generally numbered. After the dedication (folios *i*, *v*, and *ij*, *r*) there is an elaborate set of tables, including those relating to the purchase price of various grades of silver (folio *ijj*, *v*), to per cents (folio *xlix*, *r*), to the purchase price of gold (folio *lvii*, *v*), to assays (folio *lxxxj*, *r*), and to monetary affairs of various kinds. The mathematical text (folio *xcj*, *v*) consists of twenty-four pages besides the colophon (folio *cijj*, *v*). As already stated, eighteen of these pages relate chiefly to arithmetic, and six to algebra.

The book was printed in the City of Mexico in the year 1556, being the first work on mathematics to be printed outside the boundaries of Europe, except for the ancient block books of China.

In order to give some idea of the general nature of the work, a few of the problems will be set forth, chiefly those which illustrate the application of algebra as we understand the term today.

2 Typical problems not listed under algebra

1. I bought 10 varas of velvet at 20 pesos less than cost, for 34 pesos plus a vara of velvet. How much did it cost a vara?

Rule: Add 20 pesos to 34 pesos, making 54 pesos, which will be your dividend. Subtract one from 10 varas, leaving 9. Divide 54 by 9, giving 6, the price per vara.

Proof: 10 varas at 6 pesos is 60 pesos. This minus 20 pesos is 40. You paid 34 pesos plus a vara costing 6 pesos, and this gives the result, 40 pesos.

2. I bought 9 varas of velvet for as much more than 40 pesos as 13 varas at the same price is less than 70 pesos. How much did a vara cost?

Rule: Add the pesos, 40 and 70, making 110. Add the varas, 9 and 13, making 22. Dividing 110 by 22 the quotient is 5, the price of each vara.

Proof: 9 varas at 5 pesos is 45 pesos, which is 5 more than 40 pesos; and 13 varas at 5 pesos is 65, which is 5 pesos less than 70, as you see.

3. Required a number which if 8 is added to it will be a square, and if 8 is subtracted from it will also be a square.

Take half of eight, which is 4; square it, making 16; add 1, making 17, and this is the number to which if you add 8 you have 25, the root of which

is 5; and if 8 is taken from it there is left 9, the root of which is 3; for 3 times 3 is 9, as you see.

4. Find 2 numbers the sum of the squares of which will make a square number which has an integral root.

The first numbers are 3 and 4, for their squares are 9 and 16, and these added together make 25, the root of which is 5. Observe that you have 5 numbers; the first are 2 and 3; the next are 3 and 4, the proposed numbers; and there is also 5, which is their root. Place these numbers as you see in the figure below. Then use cross multiplication, saying "3 times 3 is 9, and 2 times 4 is 8". Place these numbers at the right-hand side, one under the other. Then multiply again at the top, 2 times 3 is 6; and underneath, 3 times 4 is 12. Now subtract the less from the greater, that is, 6 from 12, and there remains 6. Divide this by 5, the root of the assumed numbers, and the quotient is $1\frac{1}{5}$, one of the numbers required. Now add 8 and 9, the products of the first multiplication, and the sum is 17. Divide this by 5 and the quotient is $3\frac{2}{5}$ and this is the second required number.

Proof: The square of $1\frac{1}{5}$ is $1\frac{11}{25}$; the square of $3\frac{2}{5}$ is $11\frac{14}{25}$; and these added together, as you see, make 13.

$$\begin{array}{r} 6 \\ 5 \frac{2}{3} \frac{6}{4} \frac{3}{8} \frac{9}{8} \quad 17 \quad \frac{02}{5} \quad \frac{1}{5} \\ 12 \quad 17 \mid 3\frac{2}{5} \quad 6 \mid 1\frac{1}{5} \end{array}$$

3 Typical problems listed under algebra

Although the above problems are solved by arithmetical rules, they are essentially algebraic. Under the title *Arte Mayor* the author gives a number of examples generally involving quadratic equations, of which the following are types:

1. Find a square from which if $15\frac{3}{4}$ is subtracted the result is its own root.

Rule: Let the number be *cosa* (x). The square of half a *cosa* is equal to $\frac{1}{4}$ of a *zenso* (x^2). Adding 15 and $\frac{3}{4}$ to $\frac{1}{4}$ makes 16, of which the root is 4, and this plus $\frac{1}{2}$ is the root of the required number.

Proof: Square the square root of 16, plus half a *cosa*, which is four and a half, giving 20 and $\frac{1}{4}$, which is the square number required. From $20\frac{1}{4}$ subtract $15\frac{3}{4}$ and, and you have 4 and $\frac{1}{2}$, which is the root of the number itself.

2. A man takes passage in a ship and asks the master what he has to pay. The master says that it will not be any more than for the others. The passenger on again asking how much it would be, the master replies: "It will be the number of pesos which, multiplied by itself and added to the number, will give 1260." Required to know how much the master asked.

Rule: Let the cost be a *cosa* of pesos. Then half of a *cosa* squared makes $\frac{1}{4}$ of a *zenso*, and this added to 1260 makes 1260 and a quarter, the root of which less $\frac{1}{2}$ of a *cosa* is the number required. Reduce 1260 and $\frac{1}{4}$ to fourths; this is equal to $\frac{5041}{4}$, the root of which is 71 halves; subtract from it half of a *cosa* and there remains 70 halves, which is equal to 35 pesos, and this is what was asked for the passage.

Proof: Multiply 35 by itself and you have 1225; adding to it 35, you have 1260, the required number.

3. A man is selling goats. The number is unknown except that it is stated that a merchant asked how many there were and the seller replied: "There are so many that, the number being squared and the product quadrupled, the result will be 90,000." Required to know how many goats he had.

Afterword

Although there is much current research on medieval Indian mathematics, there are few books or articles accessible to the non-specialist. The classic history of Indian mathematics is B. Datta and A. N. Singh's *History of Hindu Mathematics* [2], but this book, written in the 1930s, does not at all reflect modern research, such as the material on power series in the three articles on India in this section. A more recent book, which has two chapters devoted to ancient and medieval Indian mathematics, is *Crest of the Peacock* [6], by George Gheverghese Joseph. We understand, however, that there are other works on Indian mathematics, now in the planning stage, which will be both comprehensive and accessible.

Since there are still massive quantities of Islamic mathematical documents still unread in libraries around the world, it is not yet possible to produce a comprehensive history of Islamic mathematics. However, two recent books provide glimpses into various aspects of this history. The first is *Episodes in the Mathematics of Medieval Islam* [1], by Len Berggren, which is designed to be comprehensible to secondary students. The second, on a somewhat higher level, is Roshdi Rashed's *The Development of Arabic Mathematics: Between Arithmetic and Algebra* [9]. This book is an organized translation of a number of articles by Rashed, originally written in French, which provide insights into the development of arithmetic and algebra in the Islamic tradition.

Probably the best general treatment of European mathematics in the medieval period is in Chapters 5 and 7 of *Science in the Middle Ages* [8], edited by David Lindberg. Many individual texts from that time period are available in English, including Leonardo of Pisa's *Book of Squares* [5] and his *Liber Abaci* [12].

A good survey of Renaissance mathematics is found in *The Italian Renaissance of Mathematics: Studies on Humanists and Mathematicians from Petrarch to Galileo* [10], by Paul Lawrence Rose. Again, many of the important texts of that time period have been translated into English. In particular, the interested reader may consult Cardano's *The Great Art, or the Rules of Algebra* [3] and see exactly how Cardano described his solution of cubic equations. Other documents from the history of the discovery of the cubic formula are available in *The History of Mathematics: A Reader* [4], edited by John Fauvel and Jeremy Gray. More information on Bombelli's work can be found in the article [7] of Federica La Nave and Barry Mazur.

Several researchers have recently been studying the *Sumario Compendioso*, trying to understand more about the background of Juan Diez. A recent article discussing this work and related works published in *New Spain* is by Edward Sandifer [11].

References

1. J. Lennart Berggren, *Episodes in the Mathematics of Medieval Islam*, Springer, New York, 1986.
2. B. Datta and A. N. Singh, *History of Hindu Mathematics*, Asia Publishing House, Bombay, 1961 (reprint of 1935–38 original).

3. Girolamo Cardano, *The Great Art, or The Rules of Algebra*, trans. and ed. by T. Richard Witmer. MIT Press, Cambridge, 1968.
4. John Fauvel and Jeremy Gray (eds.), *The History of Mathematics: A Reader*. Macmillan, London, 1987.
5. Leonardo Pisano Fibonacci, *The Book of Squares*, trans. and ed. by L. E. Sigler. Academic Press, Boston, 1987.
6. George Gheverghese Joseph, *The Crest of the Peacock: Non-European Roots of Mathematics*, 2nd edition. Princeton University Press, 2000.
7. Federica La Nave and Barry Mazur, Reading Bombelli, *The Mathematical Intelligencer* 24 (2002), 12–21.
8. David Lindberg (ed.), *Science in the Middle Ages*. University of Chicago Press, 1978.
9. Roshdi Rashed, *The Development of Arabic Mathematics: Between Arithmetic and Algebra*, Kluwer, Dordrecht, 1994.
10. Paul Lawrence Rose, *The Italian Renaissance of Mathematics: Studies on Humanists and Mathematicians from Petrarch to Galileo*, Droz, Geneva, 1975.
11. Edward Sandifer, Spanish colonial mathematics: a window on the past, *The College Mathematics Journal* 33 (2002), 266–278.
12. L. E. Sigler, *Fibonacci's Liber Abaci: A Translation into Modern English of Leonardo Pisano's Book of Calculation*, Springer, New York, 2002.



**The
Seventeenth
Century**

Foreword

The seventeenth century saw a great acceleration in the development of mathematics. In particular, it witnessed the invention of analytic geometry and the calculus, achievements accomplished through the work of numerous mathematicians. The articles in this section deal with many aspects of these important ideas. In addition, several of the articles emphasize the relationship of history to the teaching of mathematics.

The age of exploration in Europe required new and better maps. The most famous of these, produced by Gerardus Mercator in 1569, enabled sailors to plot routes of fixed compass directions as straight lines. To accomplish this, Mercator progressively increased the distances between parallels of latitude, the further they were from the equator. But Mercator himself did not explain the mathematics behind this increase in distance. In their article, Fred Rickey and Philip Tuchinsky provide this explanation, basing their work on Edward Wright's *Certain Errors in Navigation* of 1599, and relate it to the computation of the integral of the secant.

In the next article, E. A. Whitman explores the history of the cycloid. This curve, first described by Galileo, was used as a test case for the numerous new techniques being developed in the first half of the seventeenth century. Thus, Roberval found the area under the curve; Roberval, Fermat, and Descartes each found ways of drawing tangents to it; and Pascal found centers of gravity of both the region bounded by the curve and the solid formed by revolving a part of the curve around a line in the plane. Later on, it was shown that the curve was both an isochrone and a brachistochrone.

In her first article, Judith Grabiner reconsiders the purpose of Descartes' *Geometry*, finding in it a detailed guide to geometrical problem solving. She notes that to solve a problem meant, for Descartes, not only finding a curve that satisfies the conditions of the problem, but also finding its equation and the constructing the curve. In fact, much of the *Geometry* is devoted to construction techniques, especially constructions through mechanical linkages. David Dennis explores these techniques in his own article and shows how many of them can be replicated with modern geometrical software. Dennis further emphasizes Descartes' role in testing explicitly the ability of symbolic algebra to represent geometry. Today, in contrast, we usually regard the algebra as given and use it to derive geometrical properties.

James Gregory was one of the premier mathematicians of the mid-seventeenth century, but unfortunately much of his work was either neglected by his contemporaries or never published because of his untimely death. Max Dehn and E. D. Hellinger attempt to correct this injustice by discussing some of Gregory's most important ideas related to the development of calculus, including his probable knowledge of the construction of Taylor series, a half-century before Taylor's own work.

In her second article, Judith Grabiner takes us through the history of the derivative. As she notes,

the derivative was first used, then discovered, then explored, and finally defined. With well-chosen examples from the works of mathematicians from Fermat and Hudde to Cauchy and Weierstrass, she shows how the history of the ‘concept of change’ is exactly the reverse of the standard method of teaching about the derivative in a calculus course. Teachers of calculus therefore need to be aware of this history as they develop their courses.

Paul Wolfson’s article begins with Roberval’s method of finding the tangent to the cycloid, as discussed by Whitman. But he then follows Roberval in using his kinematic method to determine tangents to other curves and shows that, contrary to contemporary critics, Roberval was quite aware of the limitations and difficulties of this method. Newton used the same basic kinematic method in some of his earliest manuscripts. Wolfson enables us to follow Newton as he struggles with these ideas through a number of writings in the mid-1660s.

Another concept that Newton dealt with in his earliest writing was the power series for the logarithm. In fact, when Newton learned that Nicolas Mercator had published this series, he began to worry that some of his other discoveries on series might also be known. J. E. Hofmann discusses Mercator’s work on the logarithmic series, as well as the work of Gregory, Newton, Cotes, and others.

In the next four articles, we consider more aspects of Isaac Newton’s thought. In an article published in honor of the three-hundredth anniversary of Newton’s *Principia*, Fred Rickey takes us through the basic highlights of Newton’s life and work. We learn about the mathematical books Newton read before beginning his own work on calculus as well as aspects of Newton’s own work on the binomial theorem, optics, and gravity. Rickey also carefully dispels the myths that Newton invented the calculus so he could apply it to the study of celestial mechanics and that he originally developed the ideas in the *Principia* in algebraic form before translating them into the classical geometry. He emphasizes further that Newton’s genius was the result of his “stubborn perseverance.”

There is another ‘myth’ about Newton, one which some historians believe to be true, that in the *Principia* Newton never proved that the inverse-square law of gravitational attraction implies that a body travels in a conic path with the center of attraction at a focus. Bruce Pourciau, in the next article, shows that Newton’s argument in his masterwork, admittedly very sketchy, can be expanded into a quite rigorous proof of this theorem, one that Pourciau believes Newton had in mind.

Carl Boyer describes Newton’s use of polar coordinates, compares them with Jakob Bernoulli’s use, and then shows that Euler systematized this use in 1748. In a final article on Newton, Chris Christensen describes Newton’s method for solving ‘affected equations’ for y —that is, solving equations in two variables for one of the variables as a power series in the other. This method is based on Newton’s original method for solving polynomial equations numerically, a recursive method slightly different from what our textbooks normally call ‘Newton’s method’ today.

As the other inventor of the calculus, Leibniz also deserves some articles in this collection. In the first, by R. B. McClenon, we see how Leibniz contributed to the notion that complex numbers are a useful tool in solving equations. Leibniz showed that Cardano’s formula is valid even in the irreducible case where it produces a sum of cube roots of complex numbers. In the final selection in this part, David Dennis and Jere Confrey discuss Leibniz’s notion of a function of a curve and show how these ideas, combined with Descartes’ notion of creating a curve through mechanical linkages, can help students understand the relationship between geometry and algebra.

An Application of Geography to Mathematics: History of the Integral of the Secant

V. FREDERICK RICKEY and PHILIP M. TUCHINSKY

Mathematics Magazine 53 (1980), 162–166

Every student of the integral calculus has done battle with the formula

$$\int \sec \theta d\theta = \ln |\sec \theta + \tan \theta| + c. \quad (1)$$

This formula can be checked by differentiation or “derived” by using the substitution $u = \sec \theta + \tan \theta$, but these *ad hoc* methods do not make the formula any more understandable. Experience has taught us that this troublesome integral can be motivated by presenting its history. Perhaps our title seems twisted, but the tale to follow will show that this integral should be presented not as an application of mathematics to geography, but rather as an application of geography to mathematics.

The secant integral arose from cartography and navigation, and its evaluation was a central question of mid-seventeenth century mathematics. The first formula, discovered in 1645 before the work of Newton and Leibniz, was

$$\int \sec \theta d\theta = \ln |\tan(\theta/2 + \pi/4)| + c, \quad (2)$$

which is a trigonometric variant of (1). This was discovered, not through any mathematician’s cleverness, but by a serendipitous historical accident when mathematicians and cartographers sought to understand the Mercator map projection. To see how this happened, we must first discuss sailing and early maps so that we can explain why Mercator invented his famous map projection.

From the time of Ptolemy (c. 150 A.D.) maps were drawn on rectangular grids with one degree of latitude equal in length to one degree of longitude. When restricted to a small area, like the Mediterranean, they were accurate enough for sailors. But

in the age of exploration, the Atlantic presented vast distances and higher latitudes, and so the navigational errors due to using the “plain charts” became apparent.

The magnetic compass was in widespread use after the thirteenth century, so directions were conveniently given by distance and compass bearing. Lines of fixed compass direction were called *rhumb* lines by sailors, and in 1624 Willebrord Snell dubbed them *loxodromes*. To plan a journey one laid a straightedge on a map between origin and destination, then read off the compass bearing to follow. But rhumb lines are spirals on the globe and curves on a plain chart — facts sailors had difficulty understanding. They needed a chart where the loxodromes were represented as straight lines.

It was Gerardus Mercator (1512–1594) who solved this problem by designing a map where the lines of latitude were more widely spaced when located further from the equator. On his famous world map of 1569 ([1], p. 46), Mercator wrote:

In making this representation of the world we had ... to spread on a plane the surface of the sphere in such a way that the positions of places shall correspond on all sides with each other both in so far as true direction and distance are concerned and as concerns correct longitudes and latitudes ... With this intention we have had to employ a new proportion and a new arrangement of the meridians with reference to the parallels ... It is for these reasons that we have progressively increased the degrees of latitude towards each pole in proportion to the lengthening of the parallels with reference to the equator.

Mercator wished to map the sphere onto the plane so that both angles and distances are preserved, but he realized this was impossible. He opted for a conformal map (one which preserves angles) because, as we shall see, it guaranteed that loxodromes would appear on the map as straight lines.

Unfortunately, Mercator did not explain how he “progressively increased” the distances between parallels of latitude. Thomas Harriot (c. 1560–1621) gave a mathematical explanation in the late 1580s, but neither published his results nor influenced later work (see [6], [11]–[15]). In his *Certain Errors in Navigation* . . . [22] of 1599, Edward Wright (1561–1615) finally gave a mathematical method for constructing an accurate Mercator map. The Mercator map has its meridians of longitude placed vertically and spaced equally. The parallels of latitude are horizontal and unequally spaced. Wright’s great achievement was to show that the parallel at latitude θ should be stretched by a factor of $\sec \theta$ when drawn on the map. Let us see why.

Figure 1 represents a wedge of the earth, where AB is on the equator, C is the center of the earth, and T is the north pole. The parallel at latitude θ is a circle, with center P , that includes arc MN between the meridians AT and BT . Thus BC and NP are parallel and so angle $PNC = \theta$. The “triangles” ABC and MNP are similar figures, so

$$\frac{AB}{MN} = \frac{BC}{NP} = \frac{NC}{NP} = \sec \theta,$$

or $AB = MN \sec \theta$. Thus when MN is placed on the map it must be stretched horizontally by a factor $\sec \theta$. (This argument is not the one used by Wright [22]. His argument is two dimensional and shows

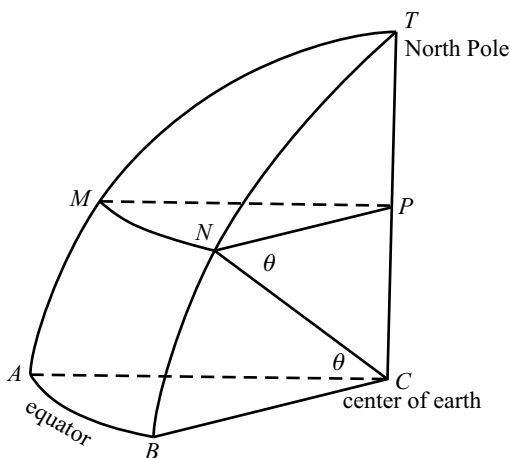


Figure 1.

that $BC = NP \sec \theta$.)

Suppose we can construct a map where angles are preserved, i.e., where the globe-to-map function is conformal. Then a loxodrome, which makes the same angle with each meridian, will appear on this map as a curve which cuts all the map’s meridians (a family of parallel straight lines) at the same angle. Since a curve that cuts a family of parallel straight lines at a fixed angle is a straight line, loxodromes on the globe will appear straight on the map. Conversely, if loxodromes are mapped to straight lines, the globe-to-map function must be conformal.

In order for angles to be preserved, the map must be stretched not only horizontally, but also vertically, by $\sec \theta$; this, however, requires an argument by infinitesimals. Let $D(\theta)$ be the distance on the map from the equator to the parallel of latitude θ , and let dD be the infinitesimal change in D resulting from an infinitesimal change $d\theta$ in θ . If we stretch vertically by $\sec \theta$, i.e., if

$$dD = \sec \theta d\theta,$$

then an infinitesimal region on the globe becomes a similar region on the map, and so angles are preserved. Conversely, if the map is to be conformal the vertical multiplier must be $\sec \theta$.

Finally, “by perpetuall addition of the Secantes”, to quote Wright, we see that the distance on the map from the equator to the parallel at latitude θ is

$$D(\theta) = \int_0^\theta \sec \theta d\theta.$$

Of course Wright did not express himself as we have here. He said ([2], pp. 312–313):

the parts of the meridian at every poynnt of latitude must needs increase with the same proportion wherewith the Secantes or hypotenusae of the arke, intercepted betweene those pointes of latitude and the aequinoctiall [equator] do increase For . . . by perpetuall addition of the Secantes answerable to the latitudes of each point or parallel vnto the summe compounded of all former secantes, . . . we may make a table which shall shew the sections and points of latitude in the meridians of the nautical planisphaere: by which sections, the parallels are to be drawne.

Wright published a table of “meridional parts” which was obtained by taking $d\theta = 1'$ and then computing the Riemann sums for latitudes below 75° . Thus

the methods of constructing Mercator’s “true chart” became available to cartographers.

Wright also offered an interesting physical model. Consider a cylinder tangent to the earth’s equator and imagine the earth to “swal [swell] like a bladder”. Then identify points on the earth with the points on the cylinder that they come into contact with. Finally unroll the cylinder; it will be a Mercator map. This model has often been misinterpreted as the cylindrical projection (where a light source at the earth’s center projects the unswollen sphere onto its tangent cylinder), but this projection is not conformal.

We have established half of our result, namely that the distance on the map from the equator to the parallel at latitude θ is given by the integral of the secant. It remains to show that it is also given by $\ln \left| \tan \left(\frac{\theta}{2} + \frac{\pi}{4} \right) \right|$.

In 1614 John Napier (1550–1617) published his work on logarithms. Wright’s authorized English translation, *A Description of the Admirable Table of Logarithms*, was published in 1616. This contained a table of logarithms of sines, something much needed by astronomers. In 1620 Edmund Gunter (1581–1626) published a table of common logarithms of tangents in his *Canon triangulorum*. In the next twenty years numerous tables of logarithmic tangents were published and so were widely available. (Not even a table of secants was available in Mercator’s day.)

In the 1640s Henry Bond (c. 1600–1678), who advertised himself as a “teacher of navigation, survey and other parts of the mathematics”, compared Wright’s table of meridional parts with a log-tan table and discovered a close agreement. This serendipitous accident led him to conjecture that

$$D(\theta) = \ln \left| \tan \left(\frac{\theta}{2} + \frac{\pi}{4} \right) \right|.$$

He published this conjecture in 1645 in Norwood’s *Epitome of Navigation*. Mainly through the correspondence of John Collins this conjecture became widely known. In fact, it became one of the outstanding open problems of the mid-seventeenth century, and was attempted by such eminent mathematicians as Collins, N. Mercator (no relation), Wilson, Oughtred, and John Wallis. It is interesting to note that young Newton was aware of it in 1665 [18], [21].

The “Learned and Industrious Nicolaus Mercator” in the very first volume of the *Philosophical Transactions of the Royal Society of London* was “willing

to lay a *Wager* against any one or more persons that have a mind to engage... *Whether the Artificial [logarithmic] Tangent-line be the true Meridian-line, yea or no?*” ([9], pp. 217–218). Nicolaus Mercator is not, as the story is often told, wagering that he knows more about logarithms than his contemporaries; rather, he is offering a prize for the solution of an open problem.

The first to prove the conjecture was, to quote Edmond Halley, “the excellent Mr. James Gregory in his *Exercitationes Geometricae*, published Anno 1668, which he did, not without a long train of Consequences and Complication of Proportions, whereby the evidence of the Demonstration is in a great measure lost, and the Reader wearied before he attain it” ([7], p. 203). Judging by Turnbull’s modern elucidation [19] of Gregory’s proof, one would have to agree with Halley. At any rate, Gregory’s proof could not be presented to today’s calculus students, and so we omit it here.

Isaac Barrow (1630–1677) in his *Geometrical Lectures* (Lect. XII, App. I) gave the first “intelligible” proof of the result, but it was couched in the geometric idiom of the day. It is especially noteworthy in that it is the earliest use of partial fractions in integration. Thus we reproduce it here in modern garb:

$$\begin{aligned} \int \sec \theta \, d\theta &= \int \frac{1}{\cos \theta} \, d\theta = \int \frac{\cos \theta}{\cos^2 \theta} \, d\theta \\ &= \int \frac{\cos \theta}{1 - \sin^2 \theta} \, d\theta \\ &= \int \frac{\cos \theta}{(1 - \sin \theta)(1 + \sin \theta)} \, d\theta \\ &= \frac{1}{2} \int \frac{\cos \theta}{1 - \sin \theta} + \frac{\cos \theta}{1 + \sin \theta} \, d\theta \\ &= \frac{1}{2} [-\ln |1 - \sin \theta| + \ln |1 + \sin \theta|] + c \\ &= \frac{1}{2} \ln |(1 + \sin \theta)/(1 - \sin \theta)| + c \\ &= \frac{1}{2} \ln |(1 + \sin \theta)^2/(1 - \sin^2 \theta)| + c \\ &= \frac{1}{2} \ln |(1 + \sin \theta)^2/\cos^2 \theta| + c \\ &= \ln |(1 + \sin \theta)/\cos \theta| + c \\ &= \ln |\sec \theta + \tan \theta| + c. \end{aligned}$$

We became interested in this topic after noting one line of historical comment in Spivak’s excellent *Calculus* (p. 326). As we ferreted out the details and shared them with our students, we found an ideal soapbox for discussing the nature of mathematics,

the process of mathematical discovery, and the role that mathematics plays in the world. We found this so useful in the classroom that we have prepared a more detailed version for our students [17].

References

The following works contain interesting information pertaining to this paper. The best concise source of information about the individuals mentioned in this paper is the excellent *Dictionary of Scientific Biography*, edited by C. C. Gillespie.

1. Anonymous, *Gerard Mercator's Map of the World (1569)*, Supplement no. 2 to *Imago Mundi*, 1961.
2. Florian Cajori, On an integration ante-dating the integral calculus, *Bibliotheca Mathematica*, 3rd series, 14 (1915) 312–319.
3. H. S. Carslaw, The story of Mercator's map. A chapter in the history of mathematics, *Math. Gazette*, 12 (1924) 1–7.
4. Georgina Dawson, Edward Wright, mathematician and hydrographer, *Amer. Neptune*, 37 (1977) 174–178.
5. Jacques Delevsky, L'invention de la projection de Mercator et les enseignements de son histoire, *Isis*, 34 (1942) 110–117.
6. Frank George, Harriot's meridional parts, *J. Inst. Navigation*, London, 21 (1968) 82–83.
7. E. Halley, An easie demonstration of the analogy of the logarithmick tangents to the meridian line or sum of secants: with various methods for computing the same to the utmost exactness, *Philos. Trans., Roy. Soc. London*, IX (1695–97) 202–214.
8. Johannes Keuning, The history of geographical map projections until 1600, *Imago Mundi*, 12 (1955) 1–24.
9. Nicolaus Mercator, Certain problems touching some points of navigation, *Philos. Trans., Roy. Soc. London*, 1 (1666) 215–218.
10. E. J. S. Parsons and W. F. Morris, Edward Wright and his work, *Imago Mundi*, 3 (1939) 61–71.
11. Jon V. Pepper, Harriot's calculation of the meridional parts as logarithmic tangents, *Archive for History of Exact Science*, 4 (1967) 359–413.
12. ———, A note on Harriot's method of obtaining meridional parts, *J. Inst. Navigation*, London, 20 (1967) 347–349.
13. ———, The study of Thomas Harriot's manuscripts, II: Harriot's unpublished papers, *History of Science*, 6 (1967) 17–40.
14. ———, Harriot's earlier work on mathematical navigation: theory and practice. With an appendix, 'The early development of the Mercator chart,' in *Thomas Harriot: Renaissance Scientist*, Clarendon Press, Oxford, 1974, John W. Shirley, editor, pp. 54–90.
15. D. H. Sadler and Eva G. R. Taylor, The doctrine of nautical triangles compendious. Part I: Thomas Harriot's manuscript (by Taylor). Part II: Calculating the meridional parts (by Sadler), *J. Inst. Navigation*, London, 6 (1953) 131–147.
16. Eva G. R. Taylor, *The Haven-Finding Art*, Hollis and Carter, London, 1971.
17. P. M. Tuchinsky, *Mercator's World Map and the Calculus*, Modules and Monographs in Undergraduate Mathematics and its Applications (UMAP) Project, Education Development Center, Newton, Mass., 1978.
18. H. W. Turnbull (ed.), *The Correspondence of Isaac Newton*, Cambridge Univ. Press, 1959–1960, vol. 1, pp. 13–16, and vol. 2, pp. 99–100.
19. ———, *James Gregory Tercentenary Memorial Volume*, G Bell & Sons, London, 1939, pp. 463–464.
20. D. W. Waters, *The Art of Navigation in England in Elizabethan and Early Stuart Times*, Yale Univ. Press, New Haven, 1958.
21. D. T. Whiteside, editor, *The Mathematical Papers of Isaac Newton*, vol. 1, Cambridge Univ. Press, 1967, pp. 466–467, 473–475.
22. Edward Wright, *Certain Errors in Navigation, Arising either of the ordinaire erroneous making or vsing of the sea Chart, Compasse, Crosse staffe, and Tables of declination of the Sunne, and fixed Starres detected and corrected*, Valentine Sims, London, 1599. Available on microfilm as part of *Early English Books 1475–1640*, reels 539 and 1018 (these two copies from 1599 have slightly different title pages). The preface and table of meridional parts have been reproduced as Origin of meridional parts, *International Hydrographic Review*, 8 (1931) 84–97.

Some Historical Notes on the Cycloid

E. A. WHITMAN

American Mathematical Monthly 50 (1943), 309–315

1 Introduction

In this paper our interest is not in a renowned mathematician, a celebrated school, or a famous problem, but in a curve, the cycloid. More particularly, our interest is to center around its relation to the mathematics of the seventeenth century, one of the great centuries in the history of the subject. This curve had the good fortune to appear at a time when mathematics was being developed very rapidly and perhaps mathematicians were fortunate that so useful a curve appeared at this time. A new and powerful tool for the study of curves was furnished by the analytic geometry, whose year of birth is commonly given as 1637. New methods for finding tangents to curves, the areas under curves, and the volumes of solids bounded by curved surfaces were being discovered at a rapid pace, and a new subject, the calculus, was in the making. In these developments the cycloid was the one curve used preeminently and nearly every mathematician of the time used it in a trial of some of his new theory, even to the extent that much of the early histories of analytic geometry, calculus, and the cycloid are closely interwoven.

In the history that follows we shall not be concerned with historical minutiae, but only with the broad outlines of the story of this curve.

2 Early history of the curve

The original discoverer of the cycloid appears to be unknown. Paul Tannery has discussed a passage by Iamblichus referring to double movement and has remarked that it is difficult to see how the cycloid could have escaped the notice of the ancients [3]. John Wallis in a letter of 1679, ascribed the discovery to Nicolas Cusanus in 1450 and also mentioned

Bouelles as one who in 1500 advanced the study of this curve. In the case of Cusanus, however, historians are agreed that Wallis was mistaken unless, says Cantor, he had access to some manuscript now lost. Now Bouelles mentions that he had observed a rolling wheel yet he seems to have considered the generated arch as a part of a circle whose radius was five-fourths that of the generating circle. The history of the cycloid becomes more definite when we come to Galileo. This scientist and teacher, famed for his telescope and microscope and as the discoverer of the isochronism of the vibrations of a pendulum, this Galileo attempted the quadrature of a cycloidal arch in 1599, at least so writes his pupil Torricelli in a publication of 1644. We here learn that Galileo had sought to measure its area and for this purpose used a balance upon which he placed a material cycloidal arch and a generating circle of like material. Always the arch was about three times as heavy as the circle, wherefore Galileo had given up his experiment since he believed that an incommensurable ratio was in question. Cantor writes of Galileo that he was the first to make this curve well known and that it was he who gave it its name. The curve was also known as a roulette and as a trochoid.

3 The work of Roberval

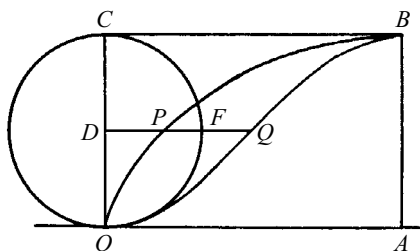
The scene now shifts to France, to the activities of Gilles Persone de Roberval, and to the problem of the quadrature of the cycloid. Going up to Paris in 1628, Roberval soon became a member of that small group of scientists and mathematicians who were wont to gather twice a week, generally at the home of Père Marin Mersenne, to discuss matters of common interest. Now Mersenne had brought the cycloid to the attention of French mathematicians at various

times and Roberval soon learned of this curve but could not immediately effect the quadrature. However, a new method of finding the areas under curves was made known in 1629 when Cavalieri submitted his notes on the theory of *indivisibles* to show his fitness for the chair of mathematics at the University of Bologna, where he was a candidate. This new theory, and its extensions later, exerted an enormous influence upon the subject of finding the areas under curves, hence on the development of the calculus. In this paper we are concerned only with one part of this theory which is known as Cavalieri's Theorem [1] and which says that if two areas are everywhere of the same width one to the other, then the areas are equal.

About 1634 Roberval effected the quadrature of the cycloid, or trochoid as he called this curve. The first publication of his proof seems to have been in 1693 when his *Traité des Indivisibles* [2] appeared. To explain the long delay in publication of this important discovery, it may be noted that the Chair of Ramus at the Collège Royale which Roberval had won in 1634, automatically became vacant every three years, to be filled again by open competition. As the incumbent set the questions it seems plausible that Roberval should conceal his methods. In this way he would have a set of questions whereby he should win the coming contests. Professor Walker states that the accident of occupying this chair caused Roberval to lose credit for many of his discoveries.

Roberval's quadrature depends upon a so-called cycloid companion curve and an application of Cavalieri's Theorem. Professor Walker [2] gives a translation of this quadrature, but we shall describe it only in a general way. This is among the very earliest of the quadratures.

Let $OABP$ be the area under the half arch of the cycloid whose generating circle has the diameter OC . Take P any point on the cycloid and take PQ equal to DF . The locus of Q will be the companion curve to the cycloid. This curve OQB is



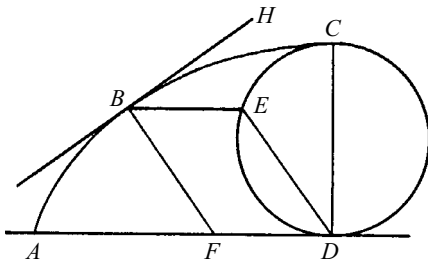
the sine curve $y = a \sin(x/a)$ where a is the radius of the generating circle, if we take the origin at the midpoint of the arc OQB , and the x -axis parallel to OA . Now by Cavalieri's Theorem, the curve OQB divides the rectangle $OACB$ into two equal parts, since to each line as DQ in $OQBC$ there corresponds an equal line in $OABQ$. The rectangle $OACB$ has its base and altitude equal respectively to the semicircumference and diameter of the generating circle, hence its area is twice that of the circle. Thus $OABQ$ has the same area as the generating circle. Also the area between the cycloid OPB and the curve OQB is equal to the area of the semicircle OFC since these two areas are everywhere of the same width one to the other. Hence the area under the half arch is one and one-half times the area of the generating circle, and the area under the arch is three times that of the generating circle.

4 Construction of the tangent

Early in 1638, Mersenne wrote to Fermat and Descartes presenting for their consideration the problem of the quadrature of the cycloid and the construction of a tangent to the curve. For a year or more previously Roberval and Fermat had been in correspondence, with Senator Carcavy as intermediary. The subjects discussed included tangents, cubatures, and centers of gravity. Mersenne's letters, however, brought to a focus the question of tangents for in August of this year Roberval, Fermat, and Descartes each gave Mersenne a method of drawing a tangent and each had a different method. In the ensuing dispute between Fermat and Descartes over the relative merits of their constructions, Roberval sided with Fermat. In turn Descartes wrote several letters to Mersenne bitterly ridiculing some of Roberval's tangent constructions which Mersenne had transmitted to him.

The question of priority in the matter of tangents we leave as unsettled and also unimportant, since each could not have borrowed from the others, so different were the methods. Part of the dispute over the relative merits of the constructions arose from different ideas as to the meaning of tangents to curves other than circles. The definition of a tangent as the limiting position of a secant had not yet been generally accepted.

We proceed to describe each of the three tangent constructions. Descartes' method is that which we now call instantaneous centers of curvature.

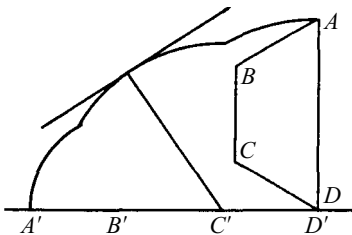


Let B be any point on the half arch of the cycloid ABC and let it be required to draw a tangent to the cycloid at B .

Draw BE parallel to the base AD cutting the circle at E . Draw BF parallel to ED and BH perpendicular to BF . BH is the required tangent.

The proof is based on the following considerations:

If a polygon $ABCD$ rolls on a straight line $A'D'$, any point A will describe a number of segments of circles whose centers will be at B', C', D' , etc. The tangents to these segments will always be perpendicular to the line joining the point of tangency to the center of the circle. Consequently if the generating circle is considered as a polygon which has an infinite number of sides, the tangent at a given point will be the line perpendicular to the line joining this point to the point where the generating circle touches the base at the same instant it passes through the point.

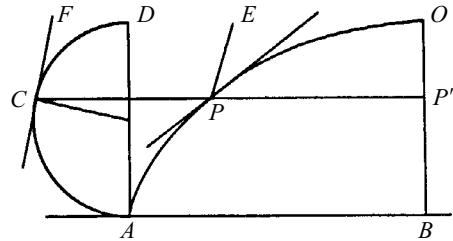


Roberval's tangent construction makes use of the composition of forces and is easily understood in connection with his particular way of stating the definition of the cycloid.

Let the diameter AD of the circle move always parallel to its original position with A on the line AB until it takes the position BO with AB equal to a semicircumference. At the same time let the point A move on the semicircle ACD in such a way that the speed of AD along AB may be equal to the speed of A along the semicircle, thus allowing A to reach the point D at the same time AD reaches BO . The point A is carried along by two motions,

its own on the semicircle and that of the diameter AD . The path of A due to these two motions is the half cycloid APO .

To construct a tangent at any point P on the cycloid, draw PP' parallel to AB cutting the semicircle at C . Then draw CF tangent to the semicircle and draw PE parallel to CF . The bisector of the angle EPP' is the required tangent since it is the resultant of two equal motions.



While finding the two components may be difficult for many curves, yet the cycloid is said to be the eleventh curve for which Roberval thus found tangents.

Fermat's construction is not unlike that of Descartes, but the proof appears to the casual reader to be quite as complicated as that of Descartes is simple. In the course of the proof a straight line is replaced by the arc of a circle. This is equivalent to assuming that an arc of a circle approaches coincidence with a certain straight line, making the method essentially one of limits. To one interested in the early approaches to the calculus, Fermat's method will be more interesting than that of Roberval or Descartes. The methods of the latter show what can be done in special cases and without the calculus. As Fermat's proof is quite long and is readily available elsewhere [4], [2], it will not be shown here.

With the area under the cycloidal arch and the tangent construction well mastered by his fellow Frenchmen, Mersenne announced these results to Galileo in 1638. Galileo, now old and blind, passed them on to his pupils Torricelli and Viviani, adding the suggestion that this curve would give a graceful form for the arch of the bridge that was projected for the nearby Arno River at Pisa. These pupils responded with a quadrature and a tangent. The interest thus kindled led Torricelli to a considerable study of the curve. In 1644 he made public his quadrature and a method of drawing a tangent. This was the earliest printed article on the cycloid.

Roberval was angered at seeing another print proofs that he considered his own discoveries. He wrote a letter to Torricelli charging plagiarism. More

specifically, Roberval charged that a certain Frenchman had written out Fermat's method of maxima and minima and Roberval's propositions on the cycloid, that these papers had come into Torricelli's hands after the death of Galileo, and that Torricelli had published them as his own. This dispute was cut short by Torricelli's early death in 1647, a death caused, according to Cajori, by this charge of plagiarism.

5 Pascal's mathematical contest

Our next episode in this history centers around Blaise Pascal, known for his *Pensées* and his *Lettres Provinciales* as well as for his mathematical works. After a brilliant early career in mathematics he had turned to theology. But suddenly the old mathematical propensity reasserted itself. Ball writes that Pascal was suffering from sleeplessness and toothache when the idea of an essay on the cycloid occurred to him. To his surprise the tooth ceased to ache. Regarding this as a divine intimation to proceed with the problem, he worked incessantly at it for eight days and completed a tolerably full account of the geometry of the cycloid. As certain questions about this curve had never been publicly answered, a prize was now offered by Pascal under the *nom de plume* of Amos Dettonville.

The year was 1658 when Newton was sixteen years old. The prizes were two in number, forty and twenty Spanish doubloons. The time allotted was June first to October first. Senator Carcavy was made recipient of the solutions offered and he, Pascal, and Roberval were the judges. The problems were as follows:

1. The area and the center of gravity of that part of a cycloidal arch above a line parallel to the base.
2. The volume and center of gravity of the volume generated when the above area is revolved about its base and also about its axis of symmetry.
3. The center of gravity of the solids formed when each body is cut by a plane parallel to its axis of revolution.

Only two contestants, Wallis and Lalouvière, had submitted offerings when time was called. Ball says that Wallis did not submit solutions for the centers of gravity, and Cajori says that Wallis made many mistakes. Both historians agree that Lalouvière was quite unequal to the task. The judges declared that neither contestant was entitled to a prize.

At the time of this contest, Sir Christopher Wren sent to Pascal his proposition on the rectification of the cycloid, not, however, including any proof. When Pascal showed this to Roberval the latter is said to have proved the proposition immediately, claiming to have known it for many years. To Wren goes the credit for the first publication and its proof [4], when Wallis published it as Wren's a year later in his *Tractatus duo*.

While the contest was on, Pascal published his *L'Histoire de la Roulette* and after the decision of the judges, his solution of the problems. With Pascal's and Wallis's publications at this time, the problems of quadrature, tangents, rectification, cubature, and centers of gravity are substantially completed in so far as the cycloid, or roulette as it was better known to the Frenchmen, was concerned. All this was accomplished in a period of about twenty-five years and before Newton's work in the calculus. The principle of *indivisibles*, or *infinites*, or whatever they had used, had in the hands of Roberval, Fermat, Torricelli, Wren, and Wallis led to important results. The cycloid curve was always being used; it was the pre-eminent curve, and its importance was to be seen later.

6 The brachistochrone problem

In another fifteen years, Huygens was using the cycloidal pendulum in an attempt to get a better chronometer and made use of the property that the evolute of the cycloid is another equal cycloid. This same Huygens discovered that a heavy particle reached the bottom of an inverted cycloidal arch in the same length of time no matter from what point on the arch it began its descent. In 1686, Leibniz wrote the equation for the curve, thus showing the rapid progress that was being made in analytic geometry. This equation is given here as Leibniz wrote it since his form shows interesting variations from those employed at present:

$$y = \sqrt{2x - xx} + \int \frac{dx}{\sqrt{2x - xx}}.$$

In the decade following the publication of this equation, the Bernoulli brothers, Jacques and Jean, published several articles on the cycloid. But we shall hurry on to one final episode in the history of the curve.

In June, 1696, Jean Bernoulli proposed a new problem which mathematicians were invited to

solve: If two points A and B are given in a vertical plane, to assign to a mobile particle M the path AMB along which, descending under its own weight, it passes from the point A to the point B in the briefest time. In later amplifying the problem Bernoulli says to choose such a curve that if the curve is replaced by a thin curve or groove and a small sphere placed in it and released, then this sphere will pass from one point to the other in the shortest possible time. Thus the famous brachistochrone problem appeared on the scene. The solution is the inverted cycloidal arch. An elaborate model of the brachistochrone formed a considerable part of the mathematics exhibit at the Golden Gate International Exposition in 1940, from which we may conclude that there is still considerable interest in the problem.

In giving out his solution [1], Jean Bernoulli wrote that a new kind of maxima and minima is required. In this solution we see that mathematics had ad-

vanced at this time as far as the calculus of variations. In a few more years there began to appear articles on general methods for determining the nature of curves formed by other rolling circles and on curves of descent under activating forces other than gravity. As the cycloid thus loses its pre-eminence, this seems a proper place to close this recital of its history.

References

1. David Smith, *Source Book in Mathematics*, pp. 644–655.
2. Evelyn Walker, *A Study of Roberval's Traité des Indivisibles*, Columbia University, 1932.
3. Paul Tannery, Pour l'histoire des lignes et surfaces courbes dans l'antiquité, *Bulletin des sciences mathématiques*, Paris, 1883, p. 284.
4. Georg Cantor, *Vorlesungen über Geschichte der Mathematik*, Volume II, 861–863, 904.

Descartes and Problem-Solving

JUDITH GRABINER

Mathematics Magazine 68 (1995), 83–97

Introduction

What does Descartes have to teach us about solving problems? At first glance it seems easy to reply. Descartes says a lot about problem-solving. So we could just quote what he says in the *Discourse on Method* [12] and in his *Rules for Direction of the Mind* ([2], pp. 9–11). Then we could illustrate these methodological rules from Descartes' major mathematical work, *La Géométrie* [13]. After all, Descartes claimed he did his mathematical work by following his “method.” And the most influential works in modern mathematics—calculus textbooks—all contain sets of rules for solving word problems, rather like this:

1. Draw a figure.
2. Identify clearly what you are trying to find.
3. Give each quantity, unknown as well as known, a name (e.g., x, y, \dots).
4. Write down all known relations between these quantities symbolically.
5. Apply various techniques to these relations until you have the unknown(s) in equations that you can solve.

The calculus texts generally owe these schemes to George Pólya's *Mathematical Discovery*, especially Chapter 2, “The Cartesian Pattern,” and Pólya himself credits them to Descartes' *Rules for Direction of the Mind* ([32], pp. 22–23, 26–28, 55–59, 129ff). So I studied those philosophical works as I began to write about Descartes and problem-solving. But the more I re-read Descartes' *Geometry*, the more convinced I became that it is from this work that his real lessons in problem-solving come. One could

claim that, just as the history of Western philosophy has been viewed as a series of footnotes to Plato, so the past 350 years of mathematics can be viewed as a series of footnotes to Descartes' *Geometry*.

Now Descartes said in the *Discourse on Method* that it didn't matter how smart you were; if you didn't go about things in the right way—with the right method—you would not discover anything. Descartes' *Geometry* certainly demonstrates a successful problem-solving method in action. Accordingly, this article will bring what historians of mathematics know about Descartes' *Geometry* to bear on the question, what can Descartes teach the mathematics community about problem-solving? To answer this question, let us look at the major types of problems addressed in the *Geometry* and at the methods Descartes used to solve them.

A first look at Descartes' *Geometry*

We have all heard that Descartes' *Geometry* contains his invention of analytic geometry. So when we look at the work, we may be quite surprised at what is *not* there. We do not see Cartesian coordinates. Nor do we see the analytic geometry of the straight line, or of the circle, or of the conic sections. In fact we do not see *any* new curve plotted from its equation. And what curves did Descartes allow? Not, as we might think, any curve that has an equation; that is secondary. He allowed only curves constructible by some mechanical device that draws them according to specified rules. Finally, we do not find the term “analytic geometry,” nor the claim that he had invented a new subject—just a new (and revolutionary) method to deal with old problems.

What we *do* see is a work that is problem-driven throughout. Descartes' *Geometry* has a purpose. It is to solve problems. Some are old, some are new; all are hard. For all the lip service in Descartes' *Discourse on Method* to mathematics as logical deduction from self-evident first principles ([12], pp. 12–13, 18–19), the *Geometry* is not like that at all. It discovers; it does not present a finished logical structure. The specific purpose of the book is to answer questions like “What is the locus of a point such that a specified condition is satisfied?” And the answer to these questions must be geometric. Not “it is such-and-such a curve,” or even “it has this equation,” but “it is this curve, it has this equation, and it can be constructed in this way.” Everything else in the *Geometry*—and that does include algebra, theory of equations, classifying curves by degree, etc.—are just means to this geometric end. To solve a problem in geometry, one must be able to construct the curve that is its solution.

The background of Descartes' *Geometry*

To appreciate how much Descartes accomplished, we must first look at some achievements of the ancient Greeks. They solved a range of locus problems, some quite complicated. To find their solutions, they too had “methods.” Greek mathematics recognized two especially useful problem-solving strategies: *reduction* and *analysis* ([25], pp. 23–24).

First, let us describe the method of reduction [in Greek, *apogōgē*]. Given a problem, we observe that we could solve it if only we could solve a second, simpler problem, and so we attack the second one instead. For instance, consider the famous problem of duplicating the cube. In modern notation, the problem is, given a^3 , to find x such that $x^3 = 2a^3$. Hippocrates of Chios showed that this problem could be reduced to the problem of finding two mean proportionals between a and $2a$. That is, again in modern notation, if we can find x and y such that:

$$a/x = x/y = y/2a, \quad (1)$$

then, eliminating y , we obtain $x^3 = 2a^3$ as required ([25], p. 23). But more geometric knowledge led to a further reduction ([25], p. 61). If we consider just the first two terms of (1),

$$a/x = x/y,$$

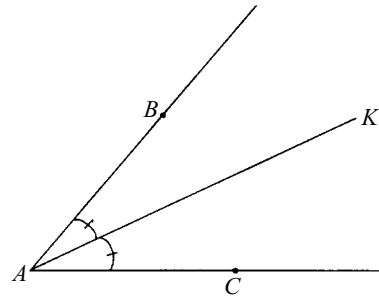


Figure 1.

we obtain $x^2 = ay$, which represents a parabola. The equation involving the first and third terms in (1) yields

$$a/x = y/2a$$

or $xy = 2a^2$, which represents a hyperbola. Thus the problem of duplicating the cube is reducible to the problem of finding the intersection of a parabola and a hyperbola. This reduction promoted Greek interest in the conic sections.

The other problem-solving strategy is what the Greeks called “analysis”—literally, “solution backwards” (*ἀνάπαλιν λύσις* [20], Vol. ii, p. 400; [25], p. 9; cp. pp. 354–360). The Greek “analysis” works like this. Suppose we want to learn how to construct an angle bisector, and suppose that we already know how to bisect a line segment. We proceed by first assuming that we have the problem solved. Then, from the assumed existence of that angle bisector, we work backward until we reach something we do know. In Figure 1, take the angle A , and draw AK bisecting it.

Then, mark off any length AB on one side of the angle, and an equal length AC on the other side. Connect B and C with the straight side BC , as in Figure 2. Now let M be the intersection of the angle bisector with the line BC . Since angle $BAM =$

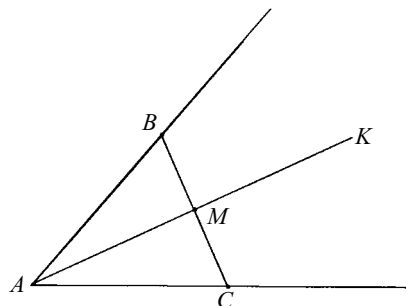


Figure 2.

angle MAC , $AB = AC$, and $AM = AM$, triangle ABM is congruent to triangle ACM . Thus M bisects BC . But wait. Recall that we already know how to bisect a line segment. Thus, we can find such an M . Now we can construct the angle bisector by reversing the process we just went through. That is, suppose we are given an angle A . To construct the angle bisector, construct $AB = AC$, construct the line BC , bisect it at M , and connect the points A and M . AM bisects the angle. This method—assuming that we have the thing we are looking for and working backwards from that assumption until we reach something we do know—was well-named “solution backwards.” Pappus of Alexandria, in the early fourth century C.E., compiled a “treasury of analysis” in which he gave the classic definition of “analysis” as “solution backwards”; described 33 works, now mostly lost, by Euclid, Apollonius, Aristaetus, and Eratosthenes, which included substantial problems solvable by the method of analysis; and provided some lemmas that illustrate problem-solving by analysis ([20], Vol. ii, pp. 399–427).

In our example of bisecting an angle, the mathematical knowledge needed was minimal. But the Greeks knew all sorts of properties of other geometric figures, notably the conic sections, and so had an extensive set of theorems to draw on in using “analysis” to solve problems in geometry ([6], pp. 21–39; [10], pp. 43–58; [20] *passim*; [25]). (The best and fullest account is that of Knorr [25].)

Thus we see that Descartes, though he championed these techniques, clearly did not invent the method of analysis and the method of reduction. Descartes’ ideas on problem-solving, moreover, have other antecedents besides the Greek mathematical tradition. First, a preoccupation with finding a universal “method” to find truth appears in the work of earlier philosophers, including the thirteenth-century Raymond Lull, whose method was to list all possible truths and select the right one, the sixteenth-century Petrus Ramus, who saw method as the key to effective teaching and to allowing learners to make their own discoveries ([29], pp. 148–149), and the seventeenth-century philosopher of science Francis Bacon, whose method to empirically discover natural laws was one of systematic induction and testing [1]. All of these seekers for method suggested that intellectual progress, unimpressive earlier in history, could be achieved once the right method for finding truth was employed. Descartes shared this view.

A second, more specific antecedent of Descartes’ work was the invention of symbolic algebra as a

problem-solving tool, a tool that was explicitly recognized as a kind of “analysis” in the Greek sense by its discoverer, Vieta, in 1591 ([6], p. 65; *cp* pp. 23, 157–173). To say “let $x =$ ” the unknown, and then calculate with x —square it, add it to itself, etc., *as if it were known*—is a powerful technique when applied to word problems both in and outside of geometry. Vieta recognized that naming the unknown and then treating it as if it were known was an example of what the Greeks called “analysis,” so he called algebra “the analytic art.” Incidentally, Vieta’s use of this term is the origin of the way we use the word “analysis” in mathematics. In the seventeenth and early eighteenth centuries, the term “analysis” was often used interchangeably with the term “algebra,” until by the mid-eighteenth century “analysis” became used for the algebra of infinite processes as opposed to that of finite ones [4].

Descartes was quite impressed with the power of symbolic algebra. But, although he had all these predecessors, Descartes combined, extended, and then exploited these earlier ideas in an unprecedented way. To see how his new method worked, we need to look at a specific problem.

Descartes’ method in action

We begin with the first important problem Descartes described solving with his new method ([13] pp. 309–314, 324–335). The problem is taken from Pappus, who said in turn that it came from Euclid and Apollonius ([13], p. 304). The problem is illustrated in Figure 3 (from [13], p. 309).

Given four lines in a plane, and given four angles. Take an arbitrary point C . Consider now the

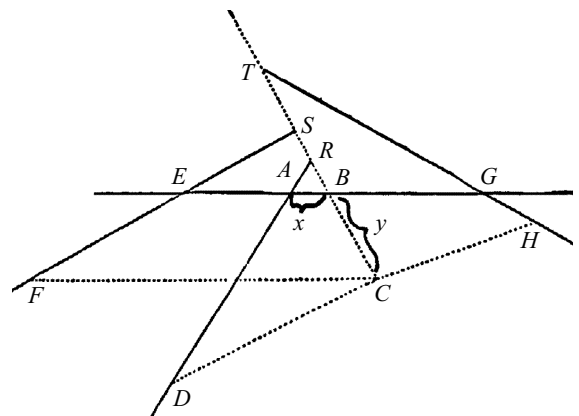


Figure 3.

distances (dotted lines) from C to the various given lines, where the distances are measured along lines making the given angles with the given lines. (For instance, the distance CD makes the given angle CDA with the given line AD .) A further condition on C is that the four distances CD , CF , CB , and CH satisfy

$$(CD \cdot CF)/(CB \cdot CH) = \text{a given constant.} \quad (2)$$

The problem is to find the locus of all such points C . For Descartes, that means to discover what curve it is, and then to construct that curve. (At this time, any reader who does not already know the answer is encouraged to conjecture what kind of curve it is — or to imagine constructing even *one* such point C .)

Here is how Descartes attacked this problem. First assume, as we must in order to draw Figure 3, that we already have one point on the curve. We will then work backwards, by the method of analysis. Draw the point C , and draw the distances. Label the distance from C to the line EG as y , and the line segment between that distance and the given line DA as x . Given these labels x and y , we use them and look for other relationships that can be derived in terms of them. For instance, independent of the choice of C , the angles in the triangle ABR are all known (since angle CBG is one of the given angles in the problem, we have angle ABR by vertical angles; angle RAB is determined by the position of the two given lines that include the segments DR and GE). Thus the shape of triangle ABR is determined, so the side RB is a fixed multiple of x . Descartes therefore called that side $(b/z) \cdot x$, where he took b/z to be a known ratio. Thus $CR = y + (b/z) \cdot x$ ([13], p. 310). Using his knowledge of geometry in this fashion, Descartes found many more such relationships, and was able to express each of the distances CD , CF , CB , and CH as a different linear function of the line segments x and y . For the case where $(CD \cdot CF)/(CB \cdot CH) = 1$, those expressions let him derive an equation between the unknowns x and y and various constants he called m , n , z , o , and p :

$$y = m - (n/z) \cdot x + \sqrt{m^2 + ox + (p/m) \cdot x^2} \quad (3)$$

([13], p. 326). Now perhaps the modern reader can guess what type of curve that equation represents. So could Descartes. From his studies of Greek geometry, Descartes knew quite a lot about the conic sections, so he said, though he did not explain, that if the coefficient of the x^2 term is zero, the points C lie on a parabola; if that coefficient is positive, on

a hyperbola; if negative, on an ellipse; etc. The positions, diameters, axes, centers, of these curves can be determined also, and he briefly discussed how to do this ([13], p. 329–332).

The reader will have observed that there is no fixed coordinate system here. Descartes labeled as x and y the lengths of line segments that arose in this particular situation. Let us also make a comment about his choice of notation. Vieta had used uppercase vowels for the unknowns, consonants for knowns. Since matters of notation are relatively arbitrary, the fact that we use Descartes' lowercase x and y , rather than Vieta's A and E , testifies to the great influence of Descartes' work on our algebra and geometry. Further, though Descartes himself wrote mm and xx rather than m^2 and x^2 ([13], p. 326), he did use raised numbers, exponents, for integer powers greater than two (e.g., [13], pp. 337, 344). Today we follow Descartes here too, using exponential notation for all powers.

The Greeks already knew that the Pappus four-line locus was a conic section. Nonetheless, the way Descartes derived this result is impressive. In line with our overall purpose, let us reflect on the method Descartes used. Why is "let x equal the first unknown" so powerful here? Because the technique of "reduction" was used by Descartes to effectively reduce a problem in geometry to a problem in algebra. Once he had done this, he could use the algorithmic power and generality of algebra to solve a formerly difficult problem with relative ease. It is an old problem-solving method, to reduce a problem to a simpler one, but because the simpler one is algebraic, Descartes had something different in kind from what had been done before. Algebra puts muscles on the problem-solving methods of analysis and reduction.

Beyond the Greeks

To fully exploit the power of algebra — to go beyond the Greeks — Descartes had to make a major break with the past. The earlier symbolic algebra of Vieta was based on the theory of geometric magnitudes inherited from the Greeks. Because of this geometric basis, the product of three magnitudes was spoken of as a volume. This created a problem: What might the product of five magnitudes be? Also, Greek geometry presupposed the Archimedean axiom: Quantities cannot be compared unless some multiple of one can exceed the other, so one cannot add a point to a line,

or an area to a solid. How then could one write x^2+x ([6], pp. 61, 84)? Descartes, like his predecessors, did not envision pure numbers, but only geometrical magnitudes. He too felt constrained to interpret all algebraic operations in geometric terms. But he invented a new geometric interpretation for algebraic equations that freed algebraists from crippling restrictions like being unable to write x^5 or x^2+x . He freed himself, and therefore freed his successors, including us. Here is how he did it.

He took a line that he called “unity,” of length one, which could be chosen arbitrarily. This let him interpret the symbol x as the area of a rectangle with one side of length x , the other of length one. He could now write x^2+x with a clear conscience, since it could be thought of as the sum of two areas. Even more important, he interpreted products as lengths of lines, so that he could interpret an arbitrary power as the length of a line. That is, the product of the line segments a and b for Descartes did not have to be the area ab , but could be another length such that $ab/a = b/1$. And the length ab could be constructed, as in Figure 4 ([13] p. 298).

In this example, the product of the lines BD and BC is constructed, given a unit line AB . Let the line segments AB and BD be laid off on the same line originating at B , and let the segment BC be laid off on a line intersecting BD . Extending BC and constructing ED parallel to AC yields the proportion $BE/BD = BC/1$, since $AB = 1$. Thus BE is the required product $BD \cdot BC$. Of course this is an easy construction, but he had to give it explicitly. Descartes’ philosophy of geometry did not let him merely assert that there was a length equal to the product of the two lines; he had to construct it. Now there was no problem in writing such expressions as x^5 . This was just the length such that $x^5/x^3 = x^2/1$.

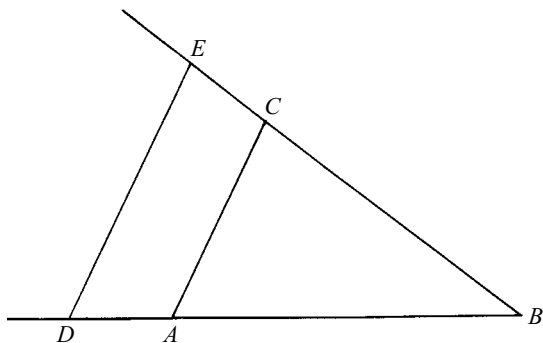


Figure 4.

By showing that all the basic algebraic operations had geometric counterparts, Descartes could use them later at will. Furthermore, he had made a major advance in writing general algebraic expressions. Because of Descartes’ innovations, later mathematicians came to consider algebra as a science of numbers, not geometric magnitudes, even though Descartes himself did not explicitly take this step. Descartes took his notational step in the service of solving geometric problems, in order to legitimize the algebraic manipulations needed to solve these geometric problems. What became a major conceptual breakthrough, then, was in the service of Descartes’ problem-solving.

Descartes could now go beyond the Greeks, extending the Pappus four-line problem to 5, 6, 12, 13, or arbitrarily many lines. With these more elaborate problems, he still followed the same method: Label line segments, work out equations. But when he found the final equation and it was not recognizable as the equation of a conic, what then? To answer this, let me give the simple example he gave, a special case of the five-line problem. He considered four parallel lines separated by a constant distance, with the fifth line perpendicular to the other four ([13], pp. 336–337). (See Figure 5.) What, he asked, is the locus of all points C such that

$$CF \cdot CD \cdot CH = CB \cdot CM \cdot AI, \quad (4)$$

where AI is the constant distance between the equally spaced parallel lines and where the distances are all measured at right angles?

Again, Descartes proceeded by analysis. Assuming that he had such a point C , he labelled the appropriate line segments x and y ($x = CM$, $y = CB$), designated the known distance AI as a , and wrote down algebraic counterparts of all known geometric relationships. For this problem they are simple ones. For instance $CD = a - y$ and $CF = a + (a - y) =$

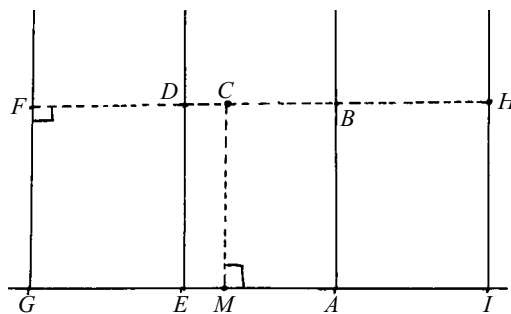


Figure 5.

$2a - y$. Thus condition (4) becomes

$$(2a - y)(a - y)(y + a) = y - x - a,$$

which, multiplied out, yields the equation ([13], p. 337)

$$y^3 - 2ay^2 - a^2y + 2a^3 = axy. \quad (5)$$

This is not a conic (it is now often called the cubical parabola of Descartes), so the next question must be, can the curve this represents be constructed? That is, given x , can we find the corresponding value of y and thus construct any point C on the curve? Until these questions are answered affirmatively, Descartes would not consider the five-line problem solved, because, for him, it is a problem in geometry. The algebraic equation was just a means to the end for Descartes; it was not in itself the solution.

So precisely what does “constructible” mean for Descartes? Can the curve represented by that cubic equation be constructed, and, if so, how?

Here another of Descartes’ methodological commitments helped him solve this problem: his commitment to generality. The ancients allowed the construction of straight lines and circles, said Descartes, but classified more complex curves as “mechanical, rather than geometrical” ([13], p. 315). Presumably this was because instruments were needed to construct them. (For instance, Nicomedes had generate the conchoid by the motion of a linkage of rulers ([25], pp. 219–220), and then used the curve in duplicating the cube and trisecting the angle.) But even the ruler and compass are machines, said Descartes, so why should one exclude other instruments ([13], p. 315; tr., p. 43)? Descartes decided to add to Euclid’s construction postulates that “two or more lines can be moved, one upon the other, determining by their intersections other curves” ([13], p. 316). The curves must be generated according to a definite rule. And for Descartes, such a rule, at least in principle, was given by the use of a mechanical device that generated a continuous motion. Exactly what this means is complex—for instance, the machine is not allowed to convert an arc length to a straight line—but Bos has provided an enlightening discussion ([3], pp. 304–322, esp. p. 314).

Figure 6 reproduces one of Descartes’ curve-constructing devices ([13], p. 320). The first curve he generated using it was produced by the intersection of moving straight lines. The straight line KN (extended as necessary) is at a fixed distance from a ruler GL . The ruler is attached to the point G , around which it can rotate. The point L can slide

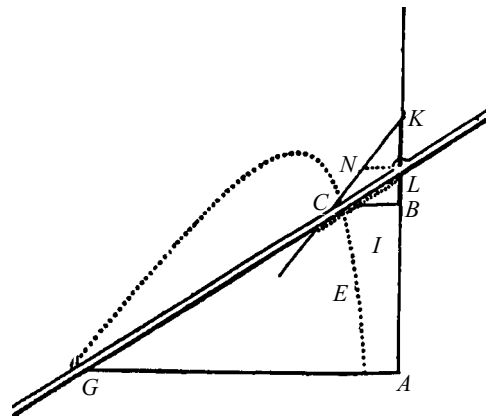


Figure 6.

along the ruler GL . The segment KL moves up the fixed line AB (extended as needed). As KL moves up, the ruler, which has a “sleeve” attached to L , rotates about G . Note that KL , KN , and the angle between them are all fixed. Then the point at which the ruler GL intersects the straight line KN extended, namely C , will be a point on the curve generated by this device.

To help the reader understand the operation of this device, I show, in Figure 7, the construction of a second point C' by this device. KL has moved up; KN thus has a new position; the ruler has rotated to a new position. Where the ruler and KN extended now intersect is another point C' on the curve. If one continues moving KL up and down, the points C , C' , etc., trace a new curve.

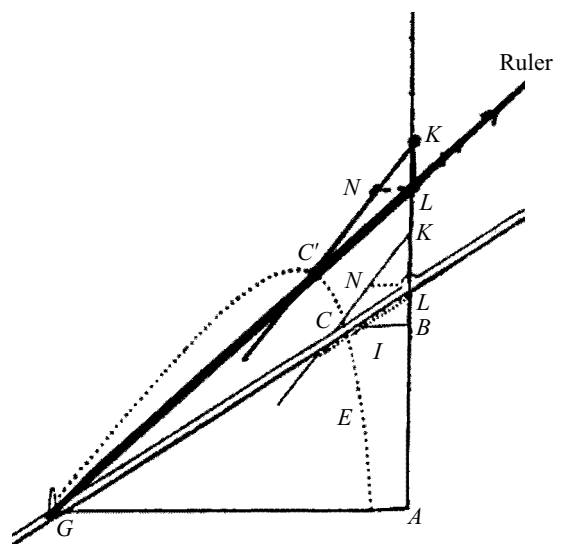


Figure 7.

But what kind of curve is it? Descartes solved this problem in his usual way. He labelled the key line segments (he let the unknowns $y = CB$ and $x = BA$, and the knowns $a = AG, b = KL$, and $c = NL$), and algebraically represented the geometric relationships between them. He then showed that if KNC is, as it is in our diagram, a straight line, the new curve generated by the points C, C' , etc., is a hyperbola ([13], p. 322). (In fact AB is one of the hyperbola's asymptotes, and the other asymptote is parallel to KN , as was shown by Jan van Schooten in his Latin edition of Descartes' *Geometry* ([13], p. 55n).) If instead of the straight line KNC , one uses a parabola whose axis is the straight line KB , the new curve constructed by the device can again be identified once its equation is found. In this case, Descartes showed by his usual method that the curve produced was precisely the cubic curve of (5) that he got for the simple five-line problem! ([13], p. 322.)

This coincidence must have suggested to Descartes that his construction method could obtain any desired curve. Also, using algebra, Descartes showed that his device would produce curves of successively higher degrees ([13], p. 321–323). For instance, when KN was a straight line, it produced a curve represented by a quadratic; when KN was a parabola, it produced a third-degree curve. Descartes, struck by the generality of these results, said that any algebraic curve could be defined as a Pappus n -line locus ([13], p. 324), but here he went too far. (For a proof that this is incorrect, see [3], pp. 332–338; incidentally, Newton was the first to try to prove that Descartes was in error on this point ([3], p. 338).) Descartes also seems to have believed that any curve with an algebraic equation could be constructed by one of his devices. And here he was right, as was shown in the nineteenth century by A. B. Kempe ([22], cited in [3], p. 324). Thus Descartes' methods really did yield results of the generality he sought. We can now understand and appreciate the claim with which Descartes' *Geometry* begins: "Every problem in geometry can easily [!] be reduced to such terms that a knowledge of the lengths of certain straight lines is sufficient for its construction." (See [13], p. 297.)

The power of Descartes' methods: tangents and equations

Descartes held that curves were admissible in geometry only if they could be constructed, but of course

he also had equations for them. Thus the study of the curves, and of many of their properties, could be advanced by the study of the corresponding equations. Let us briefly consider one example where Descartes did this.

All properties of geometric curves he had not yet discussed, he said, depend on the angles curves make with other curves ([13], pp. 341–342). This problem could be completely solved, he continued, if the normal to a curve at a given point could be found. The reader will recognize that this is an example of the reduction of one problem to another. And how does one find the normal to a curve? Again, by a reduction. It is easy to find the normal to a circle, so we can find the normal to a curve at a point by finding the normal to the circle tangent to the curve at the same point. Thus we must find such a tangent circle. And how did Descartes begin his search for that circle? By yet another reduction, this time to algebra: He sought an algebraic equation for the circle tangent to the given curve at the given point.

He did this by starting with a circle that hit the curve at two points, and then letting the two points get closer and closer together. This required, first, writing an algebraic equation for a circle that hit the curve twice. The equation for the points of intersection of that circle and the original curve would have two solutions. But "the nearer together the points . . . are taken, the less difference there is between the roots; and when the points coincide, the roots are equal"—that is, the equation has only one solution when the points coincide, and thus has only one solution when the intersecting circle becomes the tangent circle ([13], pp. 346–347). To find when the two solutions of the algebraic equation became one, Descartes in effect set the discriminant equal to zero, providing another demonstration of the power of algebraic methods to solve geometric problems. Thus, the algebraic equation let him find the tangent circle. Finally, the normal to that circle at the point of tangency gave him the normal to the curve ([6], pp. 94–95). Quite a triumph for the method of reduction!

Descartes applied this technique to find normals to several curves. For instance, he did it for the so-called ovals of Descartes ([13], pp. 360–362), whose properties, including normals, he used in optics. He also discussed finding the normal to the cubical parabola whose equation is (5) ([13], pp. 343–344). Descartes' method was the first treatment of a tangent as the limiting position of a secant to appear in print ([6], p. 95). Thus his method of normals was

a step in the direction of the calculus, as was Fermat's contemporary, independent, simpler, and more elegant method of tangents ([6], pp. 80, 94–95; [30], pp. 165–169; [5], pp. 166–169, 157–158).

There is one more important class of problems taken up in Descartes' *Geometry*, the solution of algebraic equations. As we have mentioned, classical problems like duplicating the cube required solving equations. So did constructing arbitrary points on the curve that solved a locus problem. Descartes said in fact that "all geometric problems reduce to a single type, namely the question of finding the roots of an equation." (See [13], p. 401.) Since this process was so important, if one were given an equation, it would be good to learn as much about the solutions as possible before trying to construct them geometrically.

In the last section of the *Geometry*, Descartes tried to do just this, by developing a great deal of what is now called the theory of equations. One example will suffice to illustrate his approach:

$$(x - 2)(x - 3)(x - 4)(x + 5) = 0. \quad (6)$$

Using this numerical example and multiplying it out, he obtained

$$x^4 - 4x^3 - 19x^2 + 106x - 120 = 0. \quad (7)$$

Descartes pointed out that one can see from the way the polynomial in (7) is generated from (6) that it has three positive roots and one negative one, and that the number of positive roots is given by the number of changes of sign of the coefficients (this is the principal case of what is now called Descartes' Rule of Signs). Also, a polynomial with several roots is divisible by x minus any root, and it can have as many distinct roots as its degree ([13], pp. 372–374). Descartes was not the first to have pointed out these things, but his presentation was systematic and influential, and the context made clear the importance of the results. The algebra was not an end in itself; it was all done to solve geometric problems.

The last major class of problems addressed in the *Geometry* was constructing the roots of equations of degree higher than two. Going beyond the Greek example of a cubic solved by intersecting conics, Descartes solved fifth- and sixth-degree equations. Why? They come up, he said, in geometry, if one tries to divide an angle into five equal parts ([13], pp. 412), or if one tries to solve the Pappus 12-line problem ([13], p. 324). To illustrate his solution method, he solved a sixth-degree equation with six positive roots by using intersecting cubic curves. The

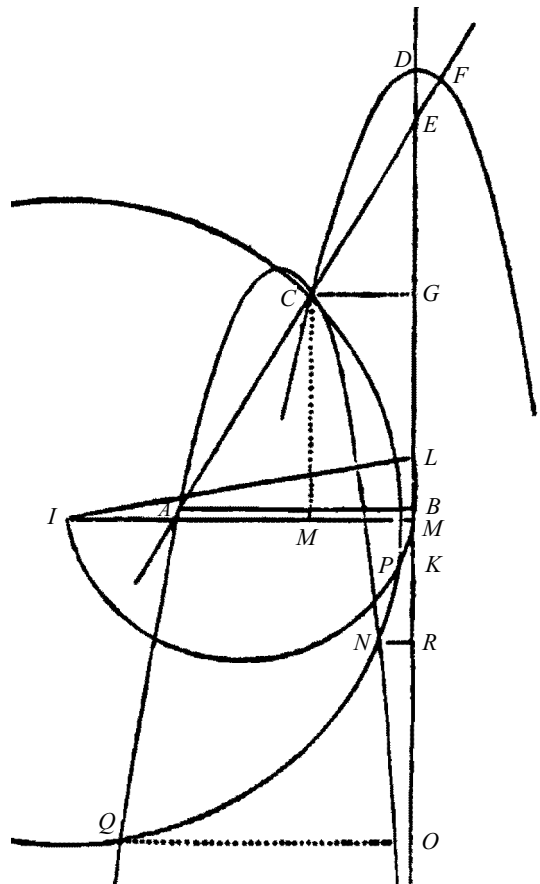


Figure 8.

curve he used was not $y = x^3$, which we might think of as simple, but the cubics he had defined as the intersections of moving conic sections and lines. In Figure 8, the diagram for one such solution is shown ([13], p. 404). The cubic curve, a portion of which is shown as NCQ , intersects the circle QNC at the points that solve the sixth-degree equation. The cubic curve involved in this construction, generated by the motion of the parabola CDF , is the cubic curve (5) once again.

Descartes said that he could construct the solution to every problem in geometry. We can now see why he thought that!

Conclusion

Now that we have seen Descartes in action, let us assess his influence on problem-solving. First, consider the mathematics that we now call "analysis." Descartes' *Geometry* solved hard problems by novel methods. There was, as an additional aid for his

successors, the simultaneous and analogous work of Fermat; though Fermat's work on analytic geometry, tangents, and areas was not printed until the 1670s, it was circulated among mathematicians in the 1630s and 1640s and exerted great influence. *Geometry* itself attracted many followers. Continental mathematicians, especially Frans van Schooten and Florimond Debeaune, wrote commentaries and added explanations for Descartes' often cryptic statements. They also extended Descartes' methods to construct other loci. The second edition of Schooten's commentary on Descartes' *Geometry* (with a Latin translation) was published in 1659–1661 together with several other influential works based on Descartes. One was Jan de Witt's *Elements of Curves*, which systematized analytic geometry, including a discussion of constructing conic sections from their equations ([6], pp. 115–116); another was Hendrik van Heuraet's work on finding arc lengths. Schooten's collection helped inspire both John Wallis and Isaac Newton. Wallis "seized upon the methods and aims of Cartesian geometry" ([6], p. 109) and went even further in replacing geometric concepts by algebraic or arithmetic ones. Many mid-seventeenth-century mathematicians, including Wallis, James Gregory, and Christopher Wren, influenced both by Descartes and by Fermat, used algebraic methods to make further progress on the problem of tangents, and — as Descartes had suggested, but did not do — to find areas. Men like van Heuraet, William Neil, and Wren also found arc lengths for some curves this way ([5], p. 162), which Descartes, who couldn't do it, had said couldn't be done ([13], p. 340). Wallis also extended the algebraic approach of Descartes to infinitesimals. In the 1660s, Isaac Newton carefully studied Schooten's edition of Descartes, using it (together with the work of men like Barrow, Wallis, and Gregory) as a key starting point in his invention of the calculus ([35], pp. 106–111, 128–130). In 1674, less than two years before his own invention of the calculus, Gottfried Wilhelm Leibniz worked his way through Descartes' *Geometry*; he was especially interested in the algebraic ideas ([21], p. 143). He later even examined some of Descartes' unpublished manuscripts ([21], pp. 182–183).

Some scholars have credited Descartes with bringing about a revolution in analysis ([7], pp. 157–159, 506; [3], p. 304; for dissenting views, see [31], pp. 110–111; [21], pp. 202–210; [18], p. 55; [19], p. 164). But at the very least we may say of the *Geometry* what Thomas Kuhn once said about Copernicus' *On the Revolution of the Celestial Orbs* ([26],

p. 134); though it may not have been revolutionary, it was "a revolution-making text." The problem-solving methods introduced in Descartes' *Geometry* and developed in the commentaries on it were clearly seminal throughout the seventeenth century, influencing both Newton and Leibniz, whether or not Descartes was the first inventor of these techniques. And such influence continued through the eighteenth century and beyond ([17], pp. 156–158, 505–507).

Incidentally, the systematic approach to analytic geometry we all learned in school is not in either Descartes or Fermat (though Fermat, unlike Descartes, did plot elementary curves from their equations), but dates from various eighteenth-century textbooks, especially those from the hands of Euler, Monge, Lagrange, and Lacroix ([16], pp. 192–224). Descartes, though, was not a textbook writer, but a problem-solver. The essence of his influence was in his new approach and his self-consciousness about method. These highlight his achievement as a problem-solver.

Second, then, let us look at his influence on problem-solving in general. The problem-solving methods we teach our students are the direct descendants of Descartes' methods. This is not because he passed them down to us in a set of rules (although he did). Nor is it because his methods work for the problems in elementary textbooks (although they do). It is because his methods solved many outstanding problems of his day. Descartes saw himself as a problem-solver because he had a method. He saw himself also as a teacher of problem-solving. One can see this even in the way he left hard questions as exercises to the reader, as he put it at the end of the *Geometry*, "to leave for others the pleasures of discovery." (See [13], p. 413.) His *Geometry* teaches us how to solve problems because it contains a set of solved problems whose successful solutions validate his methods. We may not care about the Pappus four-line problem, but we certainly prize the problem-solving power of a generalized algebra. Descartes' methods have come to us indirectly — who reads the *Geometry* nowadays? — but they have come to us because they are embedded in the work of his successors: In algebraic notation and equation theory, in analytic geometry, in calculus, in Lagrange's view that algebra is the study of general systems of operations, and in the more abstract and general subjects built upon these achievements. Because of his influence on later mathematicians, Descartes' methods are embedded also in the way we teach mathematics, in the standard collections of problems and

solutions. In fact, for routine problems, the task of applying Descartes' analytic methods is, as he intended, fairly mechanical. Some of the *Rules for Direction of the Mind* explicitly parallel the method of the *Geometry*, ([2], pp. 177–178) and Pólya is thus right to have made such rules explicit for modern students. Inventing new mathematical methods — say, like analytic geometry — is, however, not a routine task. Even here, for Descartes, “method” is crucial.

Third, then, for those of us who want to invent great and new things like analytic geometry, to teachers and students of mathematics, Descartes has something else he wants us to learn, and that is his emphasis on method in general. Here he, together with his great contemporary Sir Francis Bacon, have inspired many. For instance, Leibniz saw his differential calculus as a problem-solving method, explicitly comparing it with analytic geometry, saying “From [my differential calculus] flow all the admirable theorems and problems of this kind with such ease that there is no more need to teach and retain them than for him who knows our present algebra to memorize many theorems of ordinary geometry” ([27], excerpted in [34], p. 281). Or, in our century, there is Pólya's sophisticated emphasis on teaching about method. Let me put Descartes' lesson this way: Raise problem-solving techniques to consciousness. Reflect on the methods that are successful and on their strengths and weaknesses. Then apply them systematically in attacking new problems. That is how Descartes himself invented analytic geometry, as he said in the *Discourse on Method*: “I took the best traits of geometrical analysis and algebra, and corrected the faults of one by the other.” (See [12], pp. 13, 20.)

Fourth and last, let us briefly consider a key point in Descartes' philosophy: that the methods of mathematics could solve the problems of science. Here, Descartes the philosopher learned from Descartes the mathematician that method was important, that the right method could solve previously intractable problems. He used the ideas of reduction and analysis in his philosophy of science. For instance, he argued that all macroscopic phenomena could be explained by analyzing nature into its component parts, bits of matter in motion. (See [14], pp. 409–414 and [36], pp. 32–38.) Descartes came to believe that the most powerful methods were both general and mathematical. His *Principles of Philosophy* (1644) attempted to deduce all the laws of nature from self-evident first principles; his principles XXXVII and XXXIX

are equivalent to Newton's First Law of Motion (1687) ([8], pp. 182–183). In fact, Descartes went so far as to state that everything that could be known could be found by a method modelled on that of mathematics. He wrote,

Those long chains of reasoning, so simple and easy, which enabled the geometers to reach the most difficult demonstrations, had made me wonder whether all things knowable to man might not fall into a similar logical sequence. If so, we need only refrain from accepting as true that which is not true, and carefully follow the order necessary to deduce each one from the others, and there cannot be any propositions so abstruse that we cannot prove them, not so recondite that we cannot discover them ([12], pp. 12–13, 19).

Descartes' vision is clearly echoed by what Leibniz wrote in 1677 about his own search for a general symbolic method of finding truth: “If we could find characters or signs appropriate for expressing all our thoughts as definitely and as exactly as arithmetic expresses numbers or geometric analysis expresses lines, we could in all subjects in so far as they are amenable to reasoning accomplish what is done in Arithmetic and Geometry.” (See [28], p. 15.) Again, consider the prediction of the great prophet of progress of the Enlightenment, the Marquis de Condorcet, that Descartes' methods could solve all problems. Although the “method” of algebra “is by itself only an instrument pertaining to the science of quantities,” Condorcet wrote, “it contains within it the principles of a universal instrument, applicable to all combinations of ideas.” This could make the progress of “every subject embraced by human intelligence . . . as sure as that of mathematics.” (See [9], pp. 238, 278–279; quoted in [17], p. 222.)

Descartes has been attacked as a methodological imperialist and a reductionist and lauded as an intellectual liberator and one of the founders of modern thought (e.g., [11], [18], [24], [33]). For good or ill, the power of Descartes' vision has shaped Western thought since the seventeenth century, and his mathematical work helped inspire his philosophy. But whatever our assessment of Descartes the philosopher may be, his importance for the mathematician is clear. The history of the past 350 years of mathematics can fruitfully be viewed as the story of the triumph of Descartes' methods of problem-solving.

References

1. Bacon, Francis, *Novum Organum* [1620]. Often reprinted, e.g., in E. A. Burt, ed., *The English Philosophers from Bacon to Mill*, Modern Library, New York, 1913, pp. 24–123.
2. Beck, L. J., *The Method of Descartes: A Study of the Regulae*, Clarendon, Oxford, 1952. [Note: Descartes' *Rules for Direction of the Mind (Regulae)* were written about 1628, and published posthumously in 1701.]
3. Bos, H. J. M., On the representation of curves in Descartes' *Geométrie*, *Arch. Hist. Ex. Sci.* 1981, 295–338.
4. Boyer, Carl B., Analysis: Notes on the evolution of a subject and a name, *Math Teacher* 47 (1954), 450–462.
5. —, *The Concepts of the Calculus*, Columbia, New York, 1939.
6. —, *History of Analytic Geometry*, Scripta Mathematica, New York, 1956.
7. Cohen, I. B., *Revolution in Science*, Harvard, Cambridge, 1985.
8. —, *The Newtonian Revolution, with illustrations of the Transformation of Scientific Ideas*, Cambridge University Press, Cambridge, 1980.
9. Condorcet, Marquis de, *Sketch for a Historical Picture of the Progress of the Human Mind*, 1793, tr. J. Barraclough, in Keith Baker, ed., *Condorcet: Selected Writings*, Bobbs-Merrill, New York, 1976.
10. Coolidge, J. L., *A History of Geometrical Methods*, Oxford University Press, Oxford, 1940.
11. Davis, P. J., and Hersh, Reuben, *Descartes' Dream*, Harcourt, Brace, Jovanovich, New York, 1986.
12. Descartes, René, *Discourse on the Method of Rightly Conducting the Reason to Seek the Truth in the Sciences*, 1637, tr. L. J. Lafleur, Bobbs-Merrill, New York, 1956. The first set of page numbers in each citation in the present paper are from this translation; the second set are from the edition of Ch. Adam et P. Tannery, eds., *Oeuvres de Descartes*, Paris, 1879–1913, Vol. VI.
13. —, *The Geometry*, tr. from the French and Latin by D. E. Smith and M. L. Latham, Dover Reprint, New York, 1954. Contains a facsimile reprint of the original 1637 French edition. In here, page references from Descartes, which appear in the Dover reprint, are given from the French edition, while citations from the Smith-Latham commentary are identified by the page numbers from the Dover reprint.
14. Dijksterhuis, E. J., *The Mechanisation of the World-Picture*, tr. C. Dikshoorn, Oxford University Press, Oxford, 1961.
15. Gaukroger, Stephen, ed., *Descartes: Philosophy, Mathematics, and Physics*, Barnes and Noble, New York, 1980.
16. Gillies, Donald, ed., *Revolutions in Mathematics*, Oxford University Press, Oxford, 1992.
17. Grabiner, Judith V., The centrality of mathematics in the history of Western thought, *Math. Magazine* 61 (1988), pp. 220–230.
18. Grosholz, Emily, *Cartesian Method and the Problem of Reduction*, Clarendon Press, Oxford, 1991.
19. —, Descartes' Unification of Algebra and Geometry, in [15, pp. 156–168].
20. Heath, Thomas L., *Greek Mathematics*, 2 vols., Clarendon Press, Oxford, 1921.
21. Hofmann, J. E., *Leibniz in Paris, 1672-1676: His Growth to Mathematical Maturity*, tr. A. Prag and D. T. Whiteside, Cambridge University Press, Cambridge, 1974.
22. Kempe, A. B., On a general method of describing plane curves of the n th degree by linkwork, *Proc. Lond. Math. Soc.* 7 (1876), 213–216.
23. Klein, Jacob, *Greek Mathematical Thought and the Origin of Algebra*, The MIT Press, Cambridge, 1968.
24. Kline, Morris, *Mathematics in Western Culture*, Oxford University Press, Oxford, 1964.
25. Knorr, Wilbur, *The Ancient Tradition of Geometric Problems*, Birkhäuser, Boston, 1986.
26. Kuhn, Thomas S., *The Copernican Revolution*, Harvard, Cambridge, 1957.
27. Leibniz, G. W., De geometria recondite et analysi indivisibilium atque infinitorum, *Acta Eruditorum* 5 (1686). Excerpted in [34, pp. 281–282].
28. —, Preface to the General Science, in [37, pp. 12–17].
29. Mahoney, Michael S., The Beginnings of Algebraic Thought in the Seventeenth Century, in [15, pp. 141–155].
30. —, *The Mathematical Career of Pierre de Fermat, 1601–65*, Princeton University Press, Princeton, 1973.
31. Mancosu, Paolo, Descartes' *Geométrie* and revolutions in mathematics, in [16, pp. 83–116].
32. Pólya, George, *Mathematical Discovery: On Understanding, Learning, and Teaching Problem Solving*, John Wiley & Sons, Inc., New York, 1981.
33. Russell, Bertrand, *A History of Western Philosophy*, Simon and Schuster, New York, 1945.
34. Struik, Dirk J., *A Source Book in Mathematics, 1200–1800*, Harvard, Cambridge, 1969.
35. Westfall, Richard S., *Never At Rest: A Biography of Isaac Newton*, Cambridge University Press, Cambridge, 1980.
36. Westfall, Richard S., *The Construction of Modern Science*, John Wiley & Sons, Inc., New York, 1971.
37. Wiener P., ed. *Leibniz: Selections*, Scribner's, New York, 1951.

René Descartes' Curve-Drawing Devices: Experiments in the Relations Between Mechanical Motion and Symbolic Language

DAVID DENNIS

Mathematics Magazine 70 (1997), 163–174

1 Introduction

By the beginning of the seventeenth century it had become possible to represent a wide variety of arithmetic concepts and relationships in the newly evolved language of symbolic algebra [19]. Geometry, however, held a preeminent position as an older and far more trusted form of mathematics. Throughout the scientific revolution geometry continued to be thought of as the primary and most reliable form of mathematics, but a continuing series of investigations took place that examined the extent to which algebra and geometry might be compatible. These experiments in compatibility were quite opposite from most of the ancient classics. Euclid, for example, describes in Books 8–10 of the *Elements* a number of important theorems of number theory cloaked awkwardly in a geometrical representation¹ [16]. The experiments of the seventeenth century, conversely, probed the possibilities of representing geometrical concepts and constructions in the language of symbolic algebra. To what extent could it be done? Would contradictions emerge if one moved freely back and forth between geometric and algebraic representations?

Questions of appropriate forms of representation dominated the intellectual activities of seventeenth century Europe, not just in science and mathemat-

ics but perhaps even more pervasively in religious, political, legal, and philosophical discussions [13, 24, 25]. Seen in the context of this social history it is not surprising that mathematicians like René Descartes and G. W. von Leibniz would have seen their new symbolic mathematical representations in the context of their extensive philosophical works. Descartes' *Geometry* [11] was originally published as an appendix to his large philosophical work, the *Discourse on Method*. Conversely, political thinkers like Thomas Hobbes commented extensively on the latest developments in physics and mathematics [25, 4]. Questions of the appropriate forms of scientific symbolism and discourse were seen as closely connected to questions about the construction of the new apparatuses of the modern state. This is particularly evident, for example, in the work of the physicist Robert Boyle [25].

This paper will investigate in detail two of the curve-drawing constructions from the *Geometry* of Descartes in such a way as to highlight the issue of the coordination of multiple representations (see, e.g., [6]). The profound impact of Descartes' mathematics was rooted in the bold and fluid ways in which he shifted between geometrical and algebraic forms of representation, demonstrating the compatibility of these seemingly separate forms of expression. Descartes is touted to students today as the originator of analytic geometry, but nowhere in the *Geometry* did he ever graph an equation. Curves were constructed from geometrical actions, many of which were pictured as mechanical apparatuses. After curves had been drawn Descartes introduced

¹See, for example, Book 10, Lemma 1 before Prop. 29, where Euclid generates all Pythagorean triples geometrically even though he violates the dimensional integrity of his argument. Areas, in the form of "similar plane numbers," are multiplied by areas to yield areas. There seems to be no way to reconcile dimension and still obtain the result.

coordinates and then analyzed the curve-drawing actions in order to arrive at an equation that represented the curve. Equations did not create curves; curves gave rise to equations.² Descartes used equations to create a taxonomy of curves [20].

It can be difficult for a person well schooled in modern mathematics to enter into and appreciate the philosophical and linguistic issues involved in seventeenth century mathematics and science. We have all been thoroughly trained in algebra and calculus and have come to rely on this language and grammar as a dominant form of mathematical representation. We inherently trust that these symbolic manipulations will give results that are compatible with geometry; a trust that did not fully emerge in mathematics until the early works of Euler more than a century after Descartes. Such trust became possible because of an extensive set of representational experiments conducted throughout the seventeenth century which tested the ability of symbolic algebraic language to represent geometry faithfully [5, 7]. Descartes' *Geometry* is one of the earliest and most notable of these linguistic experiments. Because of our cultural trust in the reliability of symbolic languages applied to geometry, many of those schooled in mathematics today have learned comparatively little about geometry in its own right.

Descartes wrote for an audience with opposite predispositions. He assumed that his readers were thoroughly acquainted with geometry, in particular the works of Apollonius (ca. 200 B.C.) on conic sections [1, 15]. In order to appreciate the accomplishments of Descartes one must be able to check back and forth between representations and see that the results of symbolic algebraic manipulations are consistent with independently established geometrical results. The seventeenth century finessed an increasingly subtle and persuasive series of such linguistic experiments in the work of Roberval, Cavalieri, Pascal, Wallis, and Newton [8, 9]. These led eventually to Leibniz's creation of a general symbolic language capable of fully representing all known geometry of his day, that being his "calculus" [5, 7].

Because many of the most simple and beautiful results of Apollonius are scarcely known to modern mathematicians, it can be difficult to recreate one essential element of the linguistic achievements of Descartes — checking algebraic manipula-

tions against independently established geometrical results. In this article I will ask the reader to become a kind of intellectual Merlin and live history backwards. After we explore one of Descartes' curve-drawing devices, we will use the resulting bridge between geometry and algebra to regain a compelling result from Apollonius concerning hyperbolic tangents. The reader can choose to regard the investigation either as a philosophical demonstration of the consistency between algebra and geometry or as a simple analytical demonstration of a powerful ancient result of Apollonius. By adopting both views one gains a fully flexible cognitive feedback loop of the sort that my students and I have found most enlightening [6].

I was recently discussing my work on curve-drawing devices and their possible educational implications with a friend. His initial reaction was surprise: "Surely you don't advocate the revival of geometrical methods; progress in mathematics has been made only to the extent to which geometry has been eliminated." This claim has historical validity, especially since the eighteenth century, but my response was that such progress was possible only after mathematicians had achieved a basic faith in the ability of algebraic language to represent and model geometry accurately. I argued that one cannot appreciate the profundity of calculus unless one is aware of the issue of coordination of independent representations. Many students seem to learn and even master the manipulations of calculus without ever having questioned or tested the language's ability to model geometry precisely. Even Leibniz, no lover of geometry, would feel that such a student had missed the main point of his symbolic achievement [5]. On this point my friend and I agreed.

Descartes' curve-drawing devices poignantly raise the issue of technology and its relation to mathematical investigation. During the seventeenth century there was a distinct turning away from the classical Greek orientation that had been popular during the Renaissance in favor of pragmatic and stoic Roman philosophy. During much of the seventeenth century a class in "Geometry" would concern itself mainly with the design of fortifications, siege engines, canals, water systems, and hoisting devices — what we would call civil and mechanical engineering. Descartes' *Geometry* was not about static constructions and axiomatic proofs, but concerned itself instead with mechanical motions and their possible representation by algebraic equations. Classical problems were addressed, but they were all trans-

²Descartes' contemporary, Fermat, did begin graphing equations but his work did not have nearly the philosophical or scientific impact of Descartes'. Fermat's original problematic contexts came from financial work rather than engineering and mechanics.

formed into locus problems, through the use of a wide variety of motions and devices that went far beyond the classical restriction to straight-edge and compass. Descartes sought to build a geometry that included all curves whose construction he considered “clear and distinct” [11, 20]. An examination of his work shows that what he meant by this was any curve that could be drawn with a “linkage”, i.e., a device made of hinged rigid rods. Descartes' work indicates that he was well aware that this class of curves is exactly the class of all algebraic curves, although he gave no formal proof of this. This theorem is scarcely known among modern mathematicians, although it can be proved straightforwardly by looking at linkages that add, subtract, multiply, divide, and generate integer powers [3]. Descartes' linkage for generating any integer power was used repeatedly in the *Geometry* and has many interesting possibilities [10].

This transformation of geometry from classical static constructions to problems involving motions and their resultant loci has once again raised itself in light of modern computer technology, specifically the advent of dynamic geometry software such as *Cabri* and *Geometer's Sketchpad*. Many new educational and research possibilities have emerged recently in response to these technological developments [26]. It seems, indeed, that seventeenth century mechanical geometry may yet rise from the ashes of history and regain a new electronic life in our mathematics classrooms. (It has always had a life in our schools of engineering, where the finding of equations that model motion has always been a fundamental concern.) My own explorations of seventeenth century dynamic geometry have been conducted with a combination of physical models and devices along with computer animations made using *Geometer's Sketchpad* [18]. The first figure in this paper is taken directly from Descartes, but all the others were made using *Geometer's Sketchpad*. This software allows a more authentic historical exploration since curves are generated from geometrical actions rather than as the graphs of equations. Static figures cannot vividly convey the sense of motion that is necessary for a complete understanding of these devices. In the generation of the figures in this paper no equations were typed into the computer.

Figure 1 is reproduced from the (original) 1637 edition of Descartes' *Geometry* [11, p. 50]. Descartes described the device as follows:

Suppose the curve EC to be described by the intersection of the ruler GL and the rectilinear

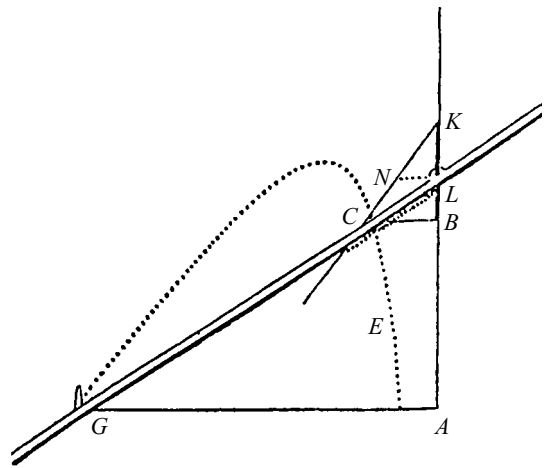


Figure 1. Descartes' Hyperbolic Device

plane figure NKL , whose side KN is produced indefinitely in the direction of C , and which, being moved in the same plane in such a way that its diameter KL always coincides with some part of the line BA (produced in both directions), imparts to the ruler GL a rotary motion about G (the ruler being hinged to the figure NKL at L). If I wish to find out to what class this curve belongs, I choose a straight line, as AB , to which to refer all its points, and on AB I choose a point A at which to begin the investigation. I say “choose this and that”, because we are free to choose what we will, for, while it is necessary to use care in the choice, in order to make the equation as short and simple as possible, yet no matter what line I should take instead of AB the curve would always prove to be of the same class, a fact easily demonstrated.

Descartes addressed here several of his main points concerning the relations between geometrical actions and their symbolic representations. His “classes of curves” refer to the use of algebraic degrees to create a taxonomy of curves. He is asserting that the algebraic degree of an equation representing a curve is independent of how one chooses to impose a coordinate system. Scale, starting point, and even the angle between axes will not change the degree of the equation, although this “fact easily demonstrated” is never given anything like a formal proof in the *Geometry*. Descartes also mentioned here the issue of a judicious choice of coordinates, an important scientific issue that goes largely unaddressed in modern mathematics curricula until an advanced

level, at which point geometry is scarcely mentioned.

Descartes went on to find the equation of the curve in Figure 1 as follows. Introduce the variables (Descartes used the term “unknown and indeterminate quantities”) $AB = y, BC = x$ (in modern notation, $C = (x, y)$), and then the constants (“known quantities”) $GA = a, KL = b$, and $NL = c$. Descartes routinely used the lower case letters x, y , and z as variables, and a, b , and c as constants; our modern convention stems from his usage. Descartes, however, had no convention about which variable was used horizontally, or in which direction (right or left) a variable was measured (here, x is measured to the left). There was, in general, no demand that x and y be measured at right angles to each other. The variables were tailored to the geometric situation. There was a very hesitant use of negative values (often called “false roots”), and in most geometric situations they were avoided.

Continuing with the derivation, since the triangles KLN and KBC are similar, we have $c/b = x/BK$, hence $BK = \frac{b}{c}x$, hence $BL = \frac{b}{c}x - b$. From this it follows that $AL = y + BL = y + \frac{b}{c}x - b$. Since triangles LBC and LAG are similar, we have $BC/BL = AG/AL$. This implies the following chain of equations:

$$\begin{aligned} \frac{x}{\frac{b}{c}x - b} &= \frac{a}{y + \frac{b}{c}x - b} \\ \Leftrightarrow x \left(y + \frac{b}{c}x - b \right) &= a \left(\frac{b}{c}x - b \right) \\ \Leftrightarrow xy + \frac{b}{c}x^2 - bx &= \frac{ab}{c}x - ab \\ \Leftrightarrow x^2 &= cx - \frac{c}{b}xy + ax - ac. \end{aligned} \quad (1)$$

Descartes left the equation in this form because he wished to emphasize its second degree. He concluded that the curve is a hyperbola. How does this follow? As we said before Descartes assumed that his readers were well acquainted with Apollonius. We will return to this issue shortly.

If one continues to let the triangle NLK rise along the vertical line, and keeps tracing the locus of the intersection of GL with NK , the lines will eventually become parallel (see Figure 2), and after that the other branch of the hyperbola will appear (see Figure 3).

These figures were made with *Geometer’s Sketchpad*, although I have altered slightly the values of the constants a, b , and c from those in Figure 1. In Figure 2, the line KN is in the asymptotic position, i.e., parallel to GL . I will hereafter refer to

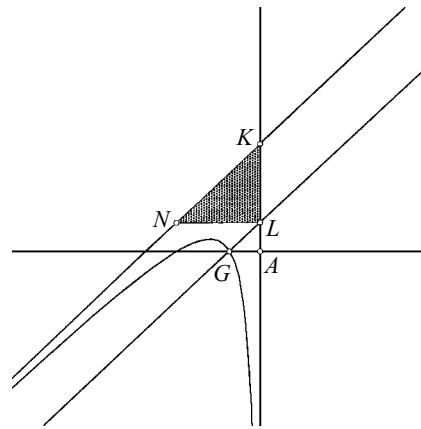


Figure 2. Descartes’ Device in the Asymptotic Position

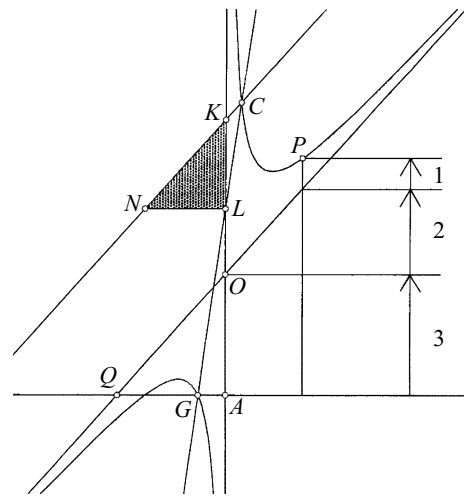


Figure 3. Geometric Display of the Terms in the Hyperbolic Equation

this particular position of the point K , as point O . In this position triangles NLK and GAL are similar, so $AK = AO = \frac{ab}{c} + b$ (the y -intercept of the asymptote). The slope of the asymptote is the same as the fixed slope of KN , i.e., b/c . (Recall that $KL = b, NL = c$, and $GA = a$.)

To rewrite Equation (1) using A as the origin in the usual modern sense, with x measured positively to the right, we can substitute $-x$ for x . With this substitution, solving Equation (1) for y yields

$$y = ab\frac{1}{x} + \frac{b}{c}x + \left(\frac{ab}{c} + b \right). \quad (2)$$

In Equation 2, the linear equation of the asymptote appears as the last two terms. In Figure 3, I have shown, to the right, the lengths that represent the values of the three terms in Equation 2, for the point

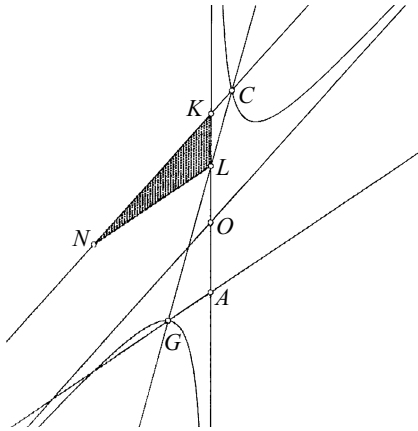


Figure 4. Hyperbola in Skewed Coordinates

P . (The labels 1, 2, and 3 represent respectively, the inverse term, the linear term, and the constant term.) Term 3 accounts for the rise from the x -axis to the level of point O (the intercept of the asymptote). Adding term 2 raises one to the level of the asymptote, and term 1 completes the ordinate to the curve.

As a geometric construction, the hyperbola is drawn from parameters that specify the angle between the asymptotes ($\angle NKL$) and a point on the curve (G). If one changes the position of the point N without changing the angle $\angle NKL$, the curve is unaffected, as in Figure 4. The derivation of the equation depends only on similarity, and not on having perpendicular coordinates. As long as GA (which determines the coordinate system) is parallel to NL , the derivation of the equation is the same except for the values of the constants $NL = c$, and $GA = a$ (both have become larger in Figure 4). Of course this equation is in the oblique coordinate system of the lines GA (x -axis) and AK (y -axis). It is the same curve geometrically, with the same form of equation, but with new constant values that refer to an oblique coordinate system. As long as angle $\angle NKL$ remains the same, and G is taken at the same distance from the line KL , the device will draw the same curve. This form of a hyperbolic equation, as an inverse term plus linear terms, depends only on using at least one of the asymptotes as an axis.

I have encountered many students who are well acquainted with the function $y = 1/x$, and yet have no idea that its graph is an hyperbola. Descartes' construction can be adjusted to draw right hyperbolas. Consider the special case in which the line KN is parallel to the x -axis (see Figure 5). The

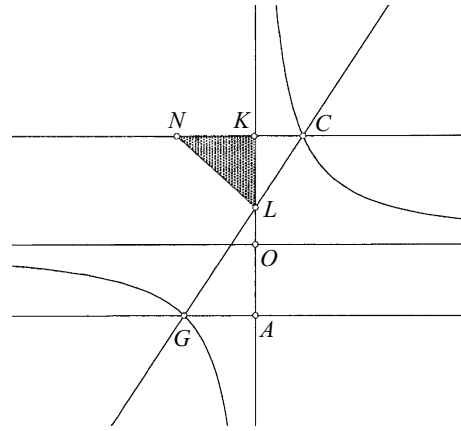


Figure 5. Device Adjusted to Draw Right Hyperbolas

point G is on the negative x -axis. Let $KC = x$, and $AK = y$ (i.e., $C = (x, y)$), $AG = a$, and $KL = b$. Now $AL = y - b$, and since triangles LKC and LAG are similar, we have $KC/KL = AG/AL$, or, equivalently

$$\frac{x}{b} = \frac{a}{y - b}.$$

Hence the curve has equation

$$y = ab\frac{1}{x} + b. \tag{3}$$

A vertical translation by b would move the origin to the point O , and letting $a = b = 1$, would put G at the vertex $(-1, -1)$, yielding the curve with equation $y = 1/x$.

Equation 3 can be seen as a special case of Equation 2, obtained by substituting ∞ for c , where c is thought of as the horizontal distance from L to the line KN . All translations and rescalings of the multiplicative inverse function can be directly seen as special members of the family of hyperbolas, using this construction.

2 Apollonius regained

How do we know that these curves are, in fact, hyperbolas? Descartes said that this is implied by Equation 1. In his commentaries on Descartes, van Schooten gives us more detail [11, p. 55, note 86]. Once again these mathematicians assumed that their readers were familiar with a variety of ratio properties from Book 2 of the *Conics* of Apollonius [1, 15] that are equivalent to Equation 1. I will not give a full set of formal proofs, but will instead suggest means for exploring these relations.

Several beautiful theorems of Apollonius concerning the relations between tangents and asymptotes

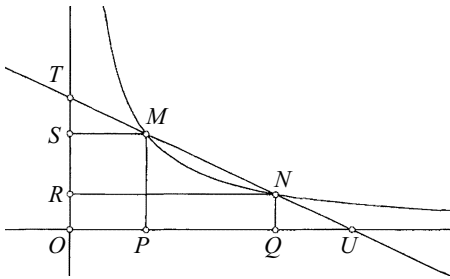


Figure 6. Hyperbola as a Family of Equal Area Triangles

are easily explored in this setting. Using the asymptotes of the curve in Figure 5 as edges to define rectangles, one sees that the points on the curve define a family of rectangles, all with the same area (see Figure 6). Indeed, if M and N are any two points on the curve, Equation 3 implies that $OPMS$ and $OQNR$ both have area equal to $a \cdot b$, the product of the constants used in drawing the curve. Another interesting geometric property is that the triangles TSM and NQU are always congruent. This congruence provides one way to dissect and compare these rectangles in a geometric manner [17].

Approaching these equations analytically, assume that the curve in Figure 6 has the equation $x \cdot y = k$ (using O as the origin). Let $M = (m, k/m)$ and $N = (n, k/n)$, i.e., $OP = m$ and $OQ = n$. The line through M and N has equation

$$y = \frac{-k}{mn}x + \left(\frac{k}{m} + \frac{k}{n}\right).$$

Hence $TO = \frac{k}{m} + \frac{k}{n}$, and, since $SO = \frac{k}{m}$, this implies that $TS = \frac{k}{n} = NQ$. Since triangles TSM and NQU are clearly similar, $TS = NQ$ implies that they are congruent and that $TM = NU$. Now let the points M and N get close to each other; then the line MN gets close to a tangent line, and one can perceive a theorem of Apollonius:

Given any tangent line to a hyperbola, the segment of the tangent contained between the two asymptotes is always bisected by the point of tangency to the curve [1, 15].

This property is a defining characteristic of hyperbolas. This simple and beautiful theorem immediately implies, among other things, that the derivative of $1/x$ is $-1/x^2$. (Look at the congruent triangles and compute the rise over run for the tangent.) This gives a student an independent geometrical check on the validity of the calculus derivation.

This bisection property of hyperbolic tangents is not restricted to the right hyperbola. Looking back

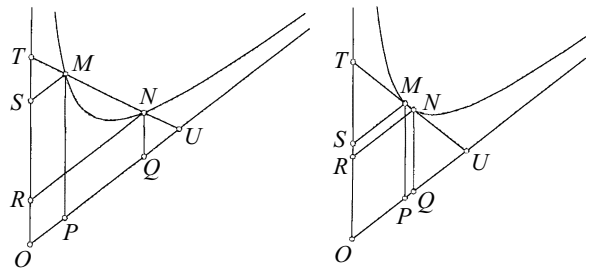


Figure 7. Bisection Property of Hyperbolic Tangents

at Figure 3 and Equation 2, one sees that any hyperbola coordinatized along both its asymptotes will always have an equation of the form $x \cdot y = k$ for some constant k . To see this, subtract off the linear and constant terms from the y -coordinate, and then rescale the x -coordinates by a constant factor that projects them in the asymptotic direction (in Figure 7 the new x -coordinate in this skew system is OQ). In the general case the curve can be seen as the set of corners of a family of equiangular parallelograms, all with the same area. In Figure 7, for any two points M and N on the curve, the parallelograms $OQNR$ and $OPMS$ have equal areas. Since the triangles TSM and NQU are congruent, by letting M and N get close together one sees that any tangent segment TU is bisected by the point of tangency (M or N).

An alternative view of the situations just described is to imagine any line parallel to TU meeting the asymptotes and the curve in corresponding points T', M' , and U' . Then the product $T'M' \cdot M'U' = TM \cdot MU$. That is to say, parallel chords between the asymptotes of a hyperbola are divided by the curve into pieces with a constant product. This follows from our discussion, because the pieces are constant projections of the sides of the parallelograms just discussed. This form of the statement was most often used by van Schooten, Newton, Euler and others in the seventeenth century. This statement (from Book 2 of Apollonius [1, 15]) was traditionally used as an identifying property of hyperbolas. This constant product was given as a proof by van Schooten that the curve drawn by Descartes' device was indeed a hyperbola [11, p. 55]. Apollonius derived these properties directly from sections of a general cone.

In this way it is possible to investigate hyperbolas, using both geometric and algebraic representations, to create a complete cognitive feedback loop. Neither representation is used as a foundation for proof; instead, one is led to a belief in a relative consis-

tency between certain aspects of geometry and algebra through checking back and forth between alternative representations. A calculus derivation of the derivative of $y = 1/x$ becomes, in this setting, a limited special case of the bisection property of hyperbolic tangents. It can be very satisfying to see symbolic algebra arrange itself into answers that are consistent with physical and geometric experience. Students of calculus can then experience the elation of Leibniz, as they build up a vocabulary of viable notation, capable of being checked against independently verifiable physical and geometric experience. Mathematical language is then seen as a powerful code for aspects of experience, rather than as the sole dictator of truth.

3 Conchoids generalized from hyperbolas

The hyperbolic device is only the beginning of what appears in Descartes' *Geometry*. He discussed several cases where curve-drawing constructions can be progressively iterated to produce curves of higher and higher algebraic degree [11, 10]. It is usually mentioned in histories of mathematics that Descartes was the first to classify curves according to the algebraic degree of their equations. This is not quite accurate. Descartes classified curves according to *pairs* of algebraic degrees; i.e., lines and conics form his first class (he used the term *genre*), curves with third or fourth degree equations form his second class, etc. [11, p. 48]. This classification is quite natural if one is working with mechanical linkages and loci. With most examples of iterated linkage, each iteration raises the degree of the curve's equation by two, with some special cases that collapse back to an odd algebraic degree [7]³. What follows is an example of such an iteration based on the hyperbolic device.

³Descartes' linkages led directly to Newton's universal method for drawing conics, which is essentially a projective method [7, 23]. This same classification by pairs of degrees is used in modern topology in the definition of "genus". The "genus" of a non-singular algebraic plane curve can be thought of topologically as the number of "handles" on the curve when defined in complex projective space. In complex projective space, linear and quadratic non-singular curves have genus 0, and are topologically sphere-like. Similarly, curves of degrees 3 and 4 are topologically torus-like, and have genus 1. Curves of degrees 5 and 6 are topologically double-holed and have genus 2, etc. In the real model, (i.e., when considering only real solutions of one real equation in 2 variables) the genus 0 curves consist of at most one oval when you join up the asymptotes. The genus 1 curves will have two ovals, which is what you'd expect when cutting through a torus by a plane, etc. (This comment was made to me by Paul Pedersen.)

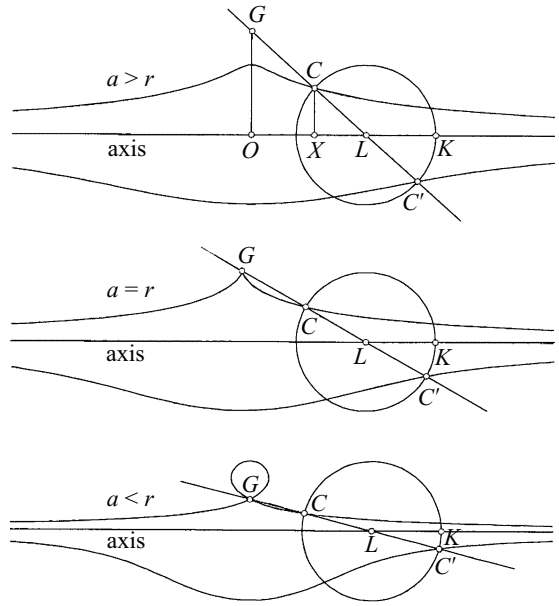


Figure 8. Conchoids Drawn by Dragging a Circle along a Line

Descartes generalized the previous hyperbola construction method by replacing the triangle KLN with any previously constructed curve. For example, let a circle with center L be moved along one axis and let the points C and C' be the intersections of the circle with the line LG , where G is any fixed point in the plane and LG is a ruler hinged at point L just as in the hyperbolic device (see Figure 8). Then C traces out a curve of degree four, known in ancient times as a *conchoid* [11, p. 55]. The two geometric parameters involved in the device are the radius of the circle (r), and the distance (a) between the point G and the axis along which L moves.

Figure 8 shows three examples of conchoids for $a > r$, $a = r$, and $a < r$. If the curve is coordinatized along the path of L , and a perpendicular line through G (OG), then its equation can be found by looking at the similar triangles GOL and CXL (top of Figure 8). Since $GO = a$, $LC = r$, $CX = y$, $OX = x$, and $XL = \sqrt{r^2 - y^2}$ one obtains the ratios of the legs of the triangles as follows:

$$\frac{\sqrt{r^2 - y^2}}{y} = \frac{\sqrt{r^2 - y^2} + x}{a}$$

This is equivalent to

$$x^2 y^2 = (r^2 - y^2)(a - y)^2,$$

an equation of fourth degree, or of Descartes' second class. (The squared form of the equation has both

branches of the curve, above and below the axis, as solutions.)

This example demonstrates Descartes' claim that, as one uses previously constructed curves to draw new curves, one gets chains of constructed curves that go up by pairs of algebraic degrees. Descartes called the conchoid a curve of the second class, i.e., of degree three or four. Dragging any rigid conic-sectioned shape along the axis, and drawing a curve in this manner will produce curves in the second class. Dragging curves of the second class will produce curves of the third class (i.e., degree five or six), etc. Descartes demonstrated this general principle through many examples [11, 7, 10], but he offered nothing like a formal proof, either geometric or algebraic. His definition of curve classes was justified by his geometric experience.

Notice that when $a \leq r$, the point G becomes a cusp or a crossover point. When singularities like cusps or crossover points occur, these tend to occur at important parts of the apparatus, like a pivot point (such as G) or a point on an axis of motion. Other important examples of this phenomena can be found in Newton's notebooks [22, 23]. I am not asserting any particular or explicit mathematical theorem here. This general observation is based upon my own historical research and empirical experience with curve-drawing devices. There are probably several ways to make this observation into an explicit mathematical statement, subject to proof (Newton attempted several [23]). There are many open questions concerning these forms of curve iteration and the relations between the parts of the physical devices and the singularities of the curves [7]. Students might benefit from such empirical experience — regardless of the extent to which they eventually formalize that experience in strictly algebraic or logical language. An instinctual sense of where curve singularities might occur is fundamentally useful in many sciences [2]. Modern computer software makes such investigations routinely possible with a minimum of technical expertise.

4 Conclusion

Descartes wrote his *Rules for the Direction of the Mind* [12] in 1625, twelve years before he would publish his famous *Geometry*. In this earlier work he emphasized the importance of making strong connections between physical actions and their possible representations in diagrams and language. Here are

a few quotes:

Rule 13: If we understand a problem perfectly, it should be considered apart from all superfluous concepts, reduced to its simplest form, and divided by enumeration into the smallest possible parts.

Rule 14: The same problem should be understood as relating to the actual extension of bodies and at the same time should be completely represented by diagrams to the imagination, for thus will it be much more distinctly perceived by the intellect.

Rule 15: It is usually helpful, also, to draw these diagrams and observe them through the external senses, so that by this means our thought can more easily remain attentive.

These lines from Descartes sound much like parts of the hands-on, problem-solving educational philosophy of mathematics put forth by the National Council of Teachers of Mathematics [21]. Descartes' entire approach to mathematics had problem solving as its foundation [14], but we must not allow ourselves to read into him too modern a perspective. He was constructing a new method of mathematical representation that responded to both the new symbolic language of his time (algebra) and to the new technology of his time (mechanical engineering). He was not seeking the broad educational goals of the NCTM. In fact, his *Geometry* was not widely read in the seventeenth century until it was republished, in 1657, with extensive commentaries by Franz van Schooten.

Nonetheless, Descartes' approach to geometry through curve-drawing devices and locus problems has important implications for education. His work connects important classical and Arabic traditions with modern algebraic formalisms [7]. It provides the missing linkages (pun intended). These linkage and loci problems, combined with the new dynamic geometry software, allow a new kind of exploration of curves that could go far towards ending the isolation of geometry in our mathematics curriculum. One can use geometrical curve generation to recreate calculus concepts such as tangents and areas in a much more elementary and physical setting [7, 8, 10], as well as to explore complicated questions about algebraic curves left open since the seventeenth century [7, 23]. Computer graphic techniques have already led to new branches of mathematics, such as fractals. Perhaps a new phase of computer-assisted empirical geometrical investigation of curves and surfaces has

already begun. If this new beginning proves as revolutionary as the century that began with Descartes' *Geometry*, then we are in for some very exciting times.

References

1. Apollonius of Perga, *On Conic Sections*, in Vol. 11 of *The Great Books of the Western World*, Encyclopedia Britannica, Chicago, IL, 1952.
2. V. I. Arnol'd, *Huygens & Barrow, Newton & Hooke*, Birkhäuser Verlag, Boston, MA, 1990.
3. I. I. Artobolevskii, *Mechanisms for the Generation of Plane Curves*, Macmillan, New York, NY, 1964.
4. F. Cajori, Controversies on mathematics between Wallis, Hobbes, and Barrow, *The Mathematics Teacher* 22 (1929), 146–151.
5. J. M. Child, *The Early Mathematical Manuscripts of Leibniz*, Open Court, Chicago, IL, 1920.
6. J. Confrey, A theory of intellectual development, *For the Learning of Mathematics* (in three consecutive issues) 14 (3) (1994), 2–8; 15 (1) (1994), 38–48; 15 (2) (1995).
7. D. Dennis, *Historical Perspectives for the Reform of Mathematics Curriculum: Geometric Curve Drawing Devices and their Role in the Transition to an Algebraic Description of Functions*, unpublished doctoral dissertation, Cornell University, Ithaca, NY, 1995.
8. D. Dennis and J. Confrey, Functions of a curve: Leibniz's original notion of functions and its meaning for the parabola, *The College Mathematics Journal* 26 (1995), 124–130.
9. D. Dennis and J. Confrey, The Creation of Continuous Exponents: A Study of the Methods and Epistemology of Alhazen and Wallis. in J. Kaput & E. Dubinsky (Eds.) *Research in Collegiate Mathematics II*, CBMS Vol. 6, pp. 33–60, American Mathematical Society, Providence, RI, 1996.
10. D. Dennis and J. Confrey, Drawing Logarithmic Curves with *Geometer's Sketchpad*: A Method Inspired by Historical Sources, in J. King and D. Schattschneider (Ed.), *Geometry Turned On: Dynamic Software in Learning, Teaching, and Research*, MAA, Washington, DC, 1997.
11. R. Descartes, *The Geometry*, Dover, New York, NY, 1954.
12. R. Descartes, *Rules For the Direction of the Mind*, Bobbs-Merrill, New York, NY, 1961.
13. M. Foucault, *The Order Of Things: An Archeology Of The Human Sciences*, Pantheon Books, New York, NY, 1970.
14. J. Grabiner, Descartes and problem solving, *Math. Magazine* 86 (1995), 83–97.
15. T. L. Heath, *Apollonius of Perga: Treatise on Conic Sections*, Barnes & Noble, New York, NY, 1961.
16. T. L. Heath, *The Thirteen Books of Euclid's Elements*, Dover, New York, NY, 1956.
17. D. Henderson, *Experiencing Geometry on Plane and Sphere*, Prentice-Hall, Englewood Cliffs, NJ, 1996.
18. N. Jackiw, *Geometer's Sketchpad* (computer program), Key Curriculum Press, Berkeley, CA, 1996.
19. J. Klein, *Greek Mathematical Thought and the Origin of Algebra*, M.I.T. Press, Cambridge, MA, 1968.
20. T. Lenoir, Descartes and the geometrization of thought: The methodological background of Descartes' geometry, *Historia Mathematica* 6 (1979), 355–379.
21. National Council of Teachers of Mathematics, *Professional Standards for Teaching Mathematics*, NCTM, Reston, VA, 1991.
22. I. Newton, *The Mathematical Papers of Isaac Newton*, Vol. 1 (1664–1666), (ed. D. T. Whiteside), Cambridge University Press, Cambridge, UK, 1967.
23. I. Newton, *The Mathematical Papers of Isaac Newton*, Vol. 2 (1667–1670), (ed. D. T. Whiteside), Cambridge University Press, Cambridge, UK, 1968.
24. B. Rotman, *Signifying Nothing: The Semiotics of Zero*, Stanford University Press, Stanford, CA, 1987.
25. S. Shapin and S. Schaffer, *Leviathan and the Air Pump: Hobbes, Boyle, and the Experimental Life*, Princeton University Press, Princeton, NJ, 1985.
26. D. Schattschneider and J. King (Eds.), *Geometry Turned On: Dynamic Software in Learning, Teaching, and Research*, MAA, Washington, DC, 1997.

Certain Mathematical Achievements of James Gregory

MAX DEHN and E. D. HELLINGER

American Mathematical Monthly 50 (1943), 149–163

For a long time the light of James Gregory did not shine as brightly as did that of John Wallis, Isaac Barrow and Isaac Newton, the other three great British mathematicians of the seventeenth century. Only recently, through the endeavors of several Scottish mathematicians, especially E. T. Whittaker, G. A. Gibson and H. W. Turnbull, Gregory's genius is revealed and fills with admiration all those interested in the development of modern mathematics.

The *James Gregory Tercentenary Memorial Volume*, edited by H. W. Turnbull [1], contains Gregory's momentous scientific correspondence, mostly with J. Collins. An extremely important supplement is the large number of Gregory's hitherto unpublished notes, recording his mathematical ideas and calculations. These notes were found in a collection of documents in the University of St. Andrews Library, written on the blank spaces of letters to Gregory. This material affords the possibility of studying his achievements and ideas.

In this paper we shall discuss Gregory's expansions of general and particular functions into series. In addition, we shall exhibit the ideas which are set forth in his first mathematical publication *Vera circuli et hyperbolae quadratura* [2]. These ideas are concerned, to some extent, with the associated problem of constructing by certain limiting processes the functions which measure the areas of circles and conics.

1 The "Taylor's series"

In a letter of February 15, 1671 to J. Collins (see *Memorial* [1], pp. 170 ff.) Gregory gives the power series for seven important functions, each with 5 or 6 terms. These functions are, if for the sake of brevity

we may use modern notations,

$$\begin{aligned} & \arctan x, \quad \tan x, \quad \sec x, \\ & \log \sec x, \quad \log \tan \left(\frac{x}{2} + \frac{\pi}{4} \right), \\ & \operatorname{arcsec}(\sqrt{2}e^x), \quad 2 \arctan \left(\tanh \frac{x}{2} \right). \end{aligned}$$

He mentions without further explanation that he had some knowledge of Newton's "universal method". Hereby, he refers to some series which Newton had discovered and which Collins had but recently communicated to him.

We may surmise that he obtained the arctangent series in a way analogous to that by which three years earlier N. Mercator [3] had found the series for $\log(1+x)$. He may have considered $\arctan x$ as the area under the curve $y = (1+x^2)^{-1}$, transformed $(1+x^2)^{-1}$ by formal division into a power series and finally integrated this infinite sum. However, there is no possibility of obtaining the other series in a similar way.

On the blank space of a letter to Gregory, dated January 29, 1671, Turnbull found a group of computations about just these seven functions [4]. The comparison of these computations with Gregory's expansions indicates the way of his thoughts. First, they include almost without exception, as many of the successive derivatives of the functions, as would be needed in finding the 5 or 6 numerical coefficients of the series by successive differentiation. Second, all coefficients in Gregory's series are correct with the exception of a single coefficient in both the expansions for $\tan x$ and for $\log \sec x$. (The second error is a consequence of the first since he obviously obtained the $\log \sec$ series by integrating the tangent

series.) Finally, all derivatives in Gregory's notes are correct with the exception of a single numerical error in the derivatives of $\tan x$, which was probably due to miscopying one number. However, using this erroneous value one finds exactly the erroneous coefficients in the series for $\tan x$ and $\log \sec x$. From these two facts, Turnbull argues conclusively that Gregory used the tables of the derivatives for the construction of his power series.

We see two possibilities for such a construction. On the one hand, we may imagine that Gregory applied in each particular case something like the "method of undetermined coefficients" together with successive differentiation. That he mentions "Newton's universal method" immediately before giving his series may be considered as supporting this assumption. In fact, if we look upon the whole of Newton's work we are justified in assuming that Gregory thought of this combined method as "Newton's universal method", even though the idea had been sketched as early as 1637 by Descartes in his *Géométrie*, and had since been applied by many other mathematicians. Nevertheless, Gregory's remark must be considered as a mere guess based upon the few results from Newton's still unpublished investigations which Collins had communicated to him with no hint about Newton's method.

On the other hand, we may suppose that Gregory could have applied the same process for an unspecified function and could have obtained the general expression for the n th coefficient of the expansion. Thus he would have anticipated Taylor's classical expansion by forty-four years. Neither the letters nor the other material, so far as published, substantiate the latter possibility. From all these facts, we may conclude that Gregory possessed a method for finding the Taylor expansion of any *particular* function, but we cannot affirm that he possessed Taylor's formula for an *unspecified* function.

It may be interesting that the second man, C. Maclaurin, whose name is closely associated with this series, deduced it seventy years later, in his *Treatise of Fluxions* (1742) by a reasoning similar to that of Gregory. Of course he applied it at once to an unspecified function. He quotes Taylor's book for the formula but could not have known Gregory's discovery then buried in the correspondence.

2 The interpolation formula

For the independent discovery by Gregory of a famous interpolation formula, full evidence is given

in a letter of his published long ago. Nevertheless, nobody seems to have realized this fact until E. T. Whittaker brought it to general notice. In the letter to Collins [5] of November 23, 1670, Gregory stated explicitly a formula which interpolates for a function $y = f(x)$ when its values at equidistant points $0, c, 2c, 3c$, are given. This formula is identical with the famous formula

$$f(x) = f(0) + \frac{x}{c} \Delta f(0) + \frac{x(x-c)}{c \cdot 2c} \Delta^2 f(0) + \frac{x(x-c)(x-2c)}{c \cdot 2c \cdot 3c} \Delta^3 f(0) + \dots, \quad (1)$$

which Newton made known some years later [6] and which mostly bears his name. It is not essential that Gregory assumes here $f(0) = 0$. Further, we may note that, of course, he did not have for the differences the notation $\Delta f(0), \Delta^2 f(0), \Delta^3 f(0), \dots$. This came into use much later under the influence of Leibniz's symbolism. He takes single letters d, f, h, \dots for these values, carefully defined by forming the sequences of the 1st, 2nd, 3rd, differences. Newton uses almost the same notation as Gregory.

In the correspondence on this formula between Collins and Gregory [7], there is mentioned the procedure which Briggs had used in extending his table of logarithms to subintervals. Briggs took differences, generalizing the older method of linear interpolation. His procedure can be considered in some way as the predecessor of the interpolation formula. However, Briggs does not state such a formula nor does he give any motivation of this procedure. Gregory's formula was given in answer to a question raised by Collins for such a motivation.

Of course, Gregory also states his formula without a proper proof, but it is obvious that he could and did verify the formula for polynomials. The same is true for Newton's first publications, although later, in the *methodus differentialis*, he sketches a way to derive the formula. It is interesting that the interpolation of tables is only *one* aim of Gregory's statement; he emphasizes strongly its use for the problem of approximate quadrature of curves and gives various formulas in this connection. Incidentally Newton [8] makes the same application of the interpolation formula.

The infinite process which is involved in this interpolation formula implies a serious mathematical difficulty which even its discoverers may have felt semiconsciously. The polynomial $P_n(x)$ of the n th degree which is given by the first $n + 1$ terms of

the formula (1) takes on the values of $f(x)$ at the equidistant points

$$0, c, 2c, \dots, nc,$$

and is determined by this property. This, obviously, is the essential fact which was discovered and communicated by Gregory and Newton. Yet they tacitly assumed that for other unspecified values of x the successive polynomials $P_n(x)$ yield an approximation to $f(x)$ which can be improved by increasing n . Apparently, they thought only of such values of x which are located *between* $0, c, \dots, nc$, that is to say, they considered only the proper problem of *interpolation*. Here the fact of the steadily improved approximation looks rather evident although a precise formulation and an exact proof were not within the range of these early developments. Things are different if one turns to the problem of *extrapolation*, considering values x *outside* the interval of the multiples of c . The published material gives no evidence that Gregory used his formula for extrapolation. And Newton in the *Philosophiae Naturalis Principia* [6] applies the interpolation formula, not in order to find the place of a comet at any time beyond the range of the observations, but only for intermediate moments.

It is important to realize this situation since the way from the interpolation formula to the Taylor series goes through a sort of extrapolation. Assuming c infinitely small, one concentrates $0, c, 2c, \dots$ in an arbitrarily small neighborhood of a fixed value and one seeks an expression for $f(x)$ at another fixed value at a finite distance. This can be done formally by applying the usual symbols of the difference and differential calculus. One has only to replace, corresponding to this limiting process, the n th difference quotient $\Delta^n y / \Delta x^n$ in Newton's formula by the n th derivative $d^n y / dx^n$. But in doing so one leaps over a very serious difficulty, using the symbols without regard to their original meaning. In fact, the higher derivatives are defined originally by iteration of the differentiation process (limit of first difference quotient) and their connection with the higher difference quotients is not trivial. And still more difficult for a critical mathematician is the whole limiting process from the interpolation formula to the infinite series. Perhaps such difficulties make us understand why Gregory did not state any connection between his two great results and why Newton, so far as we know, never formulated the Taylor series.

The first to dare to leap over these gaps was Brook Taylor in 1715 [9]. He could do so, since he obvi-

ously knew not only Newton's methods but also the concepts and notations introduced in the meantime by Leibniz. He did not use the symbols of Leibniz, but, adapting them to Newton's language, he developed a notation of his own which may, of course, appear a little awkward to us. He applied this symbolism without being influenced by the intrinsic difficulties mentioned above. Thus he came automatically from the interpolation formula to his general series by this purely formal procedure which later on was often performed unscrupulously with the help of the suggestive notation of Leibniz.

3 The binomial series

In an enclosure [10] with the letter to Collins of November 23, 1670, Gregory deals with the problem of finding the "number" of a given logarithm x ; that is to say, if we denote the base by $1 + d$, of finding $y = (1 + d)^x$. For the sake of brevity, we again use modern notations without changing anything else. Gregory gives the solution as follows:

$$(1 + d)^x = 1 + xd + \frac{x(x-1)}{1 \cdot 2}d^2 + \frac{x(x-1)(x-2)}{1 \cdot 2 \cdot 3}d^3 + \dots, \quad (2)$$

which is of course the binomial series. The comparison of Gregory's formula and notation with the statement of the interpolation theorem in the principal part of the same letter [5] shows clearly that he found his result by applying the theorem to the function $f(x) = (1 + d)^x$ using the known values at $x = 0, 1, 2, \dots$. Indeed, since the first difference of this function turns out to be

$$\begin{aligned} \Delta f(x) &= f(x+1) - f(x) \\ &= (1+d)^{x+1} - (1+d)^x = d \cdot f(x), \end{aligned} \quad (3)$$

the values of its successive differences at $x = 0$ become

$$\begin{aligned} f(0) &= 1, & \Delta f(0) &= d, \\ \Delta^2 f(0) &= d^2, & \Delta^3 f(0) &= d^3, \dots \end{aligned}$$

Thus, the interpolation formula (1) yields immediately the binomial series (2).

The correspondence of Gregory and Collins gives full evidence that this discovery of Gregory was entirely independent of Newton's investigations in the binomial theory. Gregory knew at this time only a single one of Newton's results, namely the series

for the “zone of the circle”, i.e., the series for the function

$$\int_0^x (R^2 - x^2)^{1/2} dx.$$

Collins had communicated the mere statement of the latter to him seven months previously [11]. In fact, Newton had found this series by integrating term by term the expansion of the binomial

$$(R^2 - x^2)^{1/2}.$$

Having Collins’ communication, Gregory tried hard but without success to prove the result directly. Obviously, his discovery of the general binomial theorem was in no way influenced by this knowledge and he did not guess any connection. Afterwards, he recognized suddenly that Newton’s series was a simple consequence of his own theorem and, in a letter of December 19 [12], complains much of “his own dullness”, not to have noticed the fact before. Besides, Newton’s binomial theorem did not become generally known before 1676, when, about ten years after he had found it, he communicated it to Oldenburg in the two famous letters [13] (June 6 and October 4).

It is interesting to compare the way in which Newton had discovered his theorem, as he describes it in the second of these letters, with Gregory’s deduction. We mention only the most important points, simplifying the notation as before. Newton computes first the powers $(1 + d)^n$ for the lowest integers $n = 2, 3, 4, \dots$, and discusses how to find directly the numerical coefficients of d, d^2, d^3, \dots in each of these expressions. He then makes the important remark that these coefficients in the expansion of $(1 + d)^n$ can be generated by *multiplication* of the numbers

$$\frac{n - 0}{1}, \quad \frac{n - 1}{2}, \quad \frac{n - 2}{3}, \dots,$$

that is to say, that the coefficient of d^m in the expansion of $(1 + d)^n$ is equal to

$$\frac{n(n - 1) \cdots (n - m + 1)}{1 \cdot 2 \cdots m}. \quad (4)$$

Of course, equivalent multiplicative relations for actually the same integers had been discovered a few years before by Pascal who defines them as elements of his “arithmetical triangle”, without reference to the binomials.

From this statement Newton proceeds in an extremely audacious way. He got the idea from the

procedure by which J. Wallis had developed his famous product formula for π by considering the successive integrals

$$\int_0^1 (1 - x^2)^{n/2} dx$$

for $n = 0, 1, 2, \dots$. (As a matter of fact, Newton starts in that letter with the consideration of these integrals instead of with the binomial itself.) He applies the same formula (3) also for the intermediate values $n = 1/2, 3/2, 5/2, \dots$ in order to obtain expressions for $(1 + d)^n$ with these fractional values of the exponent, although he now has to write infinite series instead of finite sums. Further generalizations enable him to state the theorem for arbitrary values of the exponent.

To be sure, neither Gregory’s nor Newton’s deduction is an exact proof in the modern sense. In some respects, Gregory’s way may seem to us more satisfactory: he deduces the result from a general theorem, the interpolation formula, and from a characteristic property of the function $(1 + d)^x$, namely the difference equation (3). On the other hand, Newton makes this almost adventurous generalization of a finite algebraic identity, deduced for integral exponents only, into an infinite series for fractional exponents. Nevertheless, there is some internal connection between the two procedures. In his investigation, Newton considers the powers of a binomial as a function of the *exponent* as does Gregory, and not as a function of the second term d of the binomial. Thus, the procedures are not so different in their essence as they are in their execution. If one compares them with the usual modern proofs of the binomial theorem, one may remark that the latter are based on the consideration of $(1 + d)^x$ as a function of d and that they use the successive *derivatives* with respect to d and the Taylor series instead of the successive *differences* with respect to x and the interpolation formula.

Newton realized the necessity of showing the way in which his consideration may be completed by a proper proof. As an example, he verifies by direct multiplication that the square of his series for $(1 + d)^{1/2}$ is equal to $1 + d$. Neither Gregory nor Newton tried to prove the convergence of the series. Such a proof was not, at this time, believed to be necessary; but certainly they had the feeling that these infinite sums determined definite numbers.

In this connection, it is interesting to find in a somewhat later letter of Gregory, dated April 9, 1672 [14], an early attempt to estimate the remainder of an

infinite series by comparing it with the geometrical series. Here, he approximates the logarithmic series

$$x + \frac{x^3}{3} + \frac{x^5}{5} + \frac{x^7}{7} + \dots$$

by expressions such as

$$x + \frac{x^3}{3} + \frac{x^5}{5} + \frac{9x^7}{7 \cdot 9 - 7 \cdot 7x^2}$$

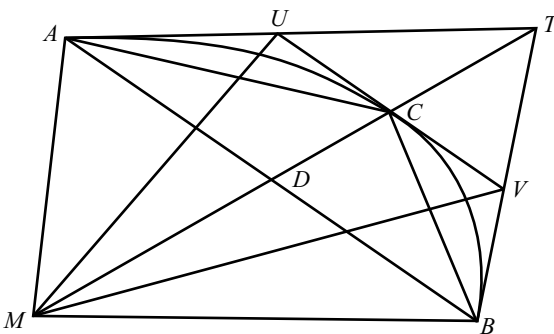
and emphasizes that the analogous expressions formed by using more terms of the original series will give a better approximation. Obviously, this estimate is obtained by comparison with the geometric series

$$x + \frac{x^3}{3} + \frac{x^5}{5} + \frac{x^7}{7} \left(1 + \frac{7x^2}{9} + \left(\frac{7x^2}{9} \right)^2 + \dots \right).$$

Thus, we see here the first step on the way which, more than a century later, led Cauchy to his convergence tests.

4 Gregory's *Vera Quadratura*

Gregory's *Vera Circuli et Hyperbolae Quadratura* [2], a small volume, contains extremely interesting and original ideas which are, to be sure, a little remote from the mathematics of his time. Even if his mathematical technique was not always sufficient to get a complete solution of the problems he saw, even if he sometimes makes incomplete deductions and wrong conclusions, the investigations show an immense creative power. He follows in some way the classical procedure of Archimedes, but reveals the algebraic content of the method. Besides, instead of calculating the perimeter of the circle as Archimedes did, he operates on areas. This enables him to deal simultaneously with the sectors of the circle, ellipse and hyperbola.



Let M be the center of a conic ACB , let AT and BT be the tangent lines at A and B , respectively, and let the straight line MT intersect AB at D and the conic at C . Gregory concludes first from fundamental properties of the conics the relations [15]:

$$AD = DB, \quad MC^2 = MD \cdot MT. \quad (5)$$

Now he draws the tangent line at C which intersects AT at U and BT at V , and compares the following pairs of polygonal areas which are inscribed in or circumscribed about the sector $MACB$: on the one hand he compares the inscribed triangle $i_0 = MAB$ with the circumscribed quadrangle $I_0 = MATB$, on the other hand the inscribed polygon $i_1 = MACB$ with the circumscribed polygon $I_1 = MAUCVB$. The polygon i_1 is composed of two equal triangles MAC and MCB ; the polygon I_1 of two equal quadrangles $MAUC$ and $MCVB$. Then, elementary properties of the conics, especially the relations (5), enable him to deduce easily two equations between these four areas as follows:

$$i_1 = \sqrt{i_0 I_0}, \quad I_1 = \frac{2i_1 I_0}{i_1 + I_0}.$$

Now, operating on the triangles MAC and MCB , and on the quadrangles $MAUC$ and $MCVB$ in the same way as he had operated on the triangle MAB and the quadrangle $MATB$, he gets four triangles of equal areas $i_2/4$, inscribed in the sector $MACB$, and four quadrangles of equal areas $I_2/4$ circumscribed about the same sector. Obviously, he obtains:

$$i_2 = \sqrt{i_1 I_1}, \quad I_2 = \frac{2i_2 I_1}{i_2 + I_1}.$$

Repeating the same operation n times, he constructs for each successive $n = 3, 4, \dots$ an inscribed polygonal area i_n composed of 2^n equal triangles, and a circumscribed one I_n , composed of 2^n equal quadrangles. The successive areas are given by:

$$i_{n+1} = \sqrt{i_n I_n}, \quad I_{n+1} = \frac{2i_{n+1} I_n}{i_{n+1} + I_n} = \frac{2i_n I_n}{i_n + \sqrt{i_n I_n}} \quad (n = 0, 1, 2, \dots). \quad (6)$$

Geometrically it is obvious that the area S of the sector $MACB$ lies between each pair i_n, I_n , and that, if n increases indefinitely, these areas will approach S as closely as one desires, one sequence increasing from below, the other decreasing from above. But Gregory is not satisfied with this visual

evidence. He recognizes in the successive construction of the i_n, I_n a new arithmetic operation which yields the value S , and therefore he feels a necessity to *prove* what we call the convergence of the limiting processes

$$\lim_{n \rightarrow \infty} i_n = \lim_{n \rightarrow \infty} I_n = S. \quad (7)$$

In fact, with that high degree of exactness which we find in the classical Greek mathematics, he first shows that

$$|I_{n+1} - i_{n+1}| < \frac{1}{2}|I_n - i_n|$$

and then concludes that $|I_n - i_n|$ becomes smaller than any given number if n is sufficiently large.

To realize the mathematical importance of Gregory's method we may state that, for the circle and ellipse where $I_0 > i_0$, the area S can be expressed as follows:

$$S = I_0 \sqrt{\frac{i_0}{I_0 - i_0}} \arctan \sqrt{\frac{I_0 - i_0}{i_0}}. \quad (8)$$

For the circle, the first factor is simply $\frac{1}{2}MA^2$, the second the angle $\theta = BMA$. For the hyperbola where $I_0 < i_0$, we have only to interchange I_0 and i_0 and to replace the arctangent function by the inverse of the hyperbolic tangent function. If we use imaginary numbers, we recognize that we have the same analytic function, since $\tanh ix = i \tan x$. But Gregory has discovered, without applying imaginary numbers, that the same analytical process — the approximation by the formulas (7), (8) — yields the area of the hyperbola as well as the area of the ellipse. In other words, he has found, for the first time in history, the analytical connection between the quadrature of sectors of the ellipse (or of the circle) and the quadrature of sectors of the hyperbola.

The history of these quadratures is interesting. We may assume that astronomical practice originally suggested the introduction of the arc of a circle as independent variable and the coordinates of the point on the circumference as dependent variables, that is to say, the introduction of the circular functions sine, cosine, and so on. This development may be connected with the fact that Archimedes investigated primarily the rectification of the circle instead of the quadrature. But the rectification of the general conics is an entirely different and much more difficult problem. In considering the *area* of the circular sectors Gregory was able to find one single analytical process for the quadrature of all conics.

Now, it has been known since the middle of the 17th century that the quadrature of the hyperbola is connected with the logarithmic function. Therefore, it was obvious to Gregory himself that he had found *one* analytical process for getting from algebraic expressions to logarithmic functions as well as to inverses of the circular functions.

This discovery is generally ascribed to Euler who, some seventy years later, arrived at the connection between the exponential function and the circular functions by using formal operations in the field of complex numbers. It is doubtful whether Euler considered hyperbolic functions as analogous to circular functions and whether he used, in this respect, the analytical analogy between the processes of quadrature of circular and hyperbolic sectors.

The comparison of Euler's and Gregory's achievements may enhance our admiration for Gregory's genius. Indeed, it is not easy to connect in the field of real numbers the two integrals

$$\int \sqrt{1 - x^2} dx \quad \text{and} \quad \int \sqrt{1 + x^2} dx$$

or

$$\int \frac{1}{\sqrt{1 + x^2}} dx \quad \text{and} \quad \int \frac{1}{\sqrt{1 - x^2}} dx.$$

As we have seen, this was achieved by Gregory.

In his *Appendicula ad veram circuli et hyperbolae quadraturam* of 1668 [16] Gregory gives an array of linear combinations of the first i_n and I_n with definite numerical coefficients which yield much better approximations to the area S than do i_n and I_n themselves. Gregory was extremely offended that Huygens did not acknowledge his work to be an essential improvement over his older methods. Therefore he tried to make obvious the strength of the new theory by stating numerous new and surprising results without revealing how he had found them. Turnbull [17] has verified that, for the circle, one gets exactly Gregory's approximations if one first expresses i_n and I_n in terms of trigonometric functions of the angle θ , then expands these expressions in power series in θ , and finally forms such linear combinations of them which begin with the term θ and contain afterwards as many vanishing coefficients as possible. Analogous considerations are valid for the hyperbola. If Gregory operated in this manner he must have known the first terms of the power series for trigonometric and hyperbolic functions as early as 1668. Indeed, it is possible that he got this knowledge without using differentiation, but the published

material does not seem to contain anything to corroborate this.

There are two other points in Gregory's speculations which particularly reveal the range of his mathematical ideas with respect to the actual later development of our science. First, the recurrent construction of the areas i_n, I_n is with him only *one* example of a very general, new analytic process which he coordinates as the "sixth" operation along with the five traditional operations (addition, subtraction, multiplication, division, and extraction of roots). In the introduction, he proudly states "*ut haec nostra inventio addat arithmeticae aliam operationem et geometriae aliam rationis speciem, ante incognitam orbi geometrico.*" This operation is, as a matter of fact, our modern limiting process. Clearly, his idea is, if we formulate it in modern language without changing the notions, to investigate two sequences of quantities a_1, a_2, \dots and b_1, b_2, \dots defined by the recurrent equations

$$a_{n+1} = \phi(a_n, b_n), \quad b_{n+1} = \chi(a_n, b_n) \quad (9)$$

$$n = 1, 2, 3, \dots$$

He uses the word "convergent" for these sequences, very probably for the first time in history, if for each n

$$0 < b_{n+1} - a_{n+1} < b_n - a_n.$$

Of course, this definition does not conform completely to our precise notion of convergence; but in applying his notion he proves in most cases the correct and sufficient inequality

$$0 < b_{n+1} - a_{n+1} < \rho(b_n - a_n)$$

where $\rho < 1$ is independent of n . (In his original problem, he has, as seen previously, $\rho = \frac{1}{2}$.) Then he concludes that the "last convergent terms" of the sequences a_n and b_n are equal, and he calls them *terminatio* of the sequences. In his original problem this *terminatio* is the area S .

From his further examples we may mention the following ones:

$$a_{n+1} = a_n + \alpha(b_n - a_n),$$

$$b_{n+1} = b_n + \beta(b_n - a_n) \quad (10)$$

and

$$a_{n+1} = \frac{2a_nb_n}{a_n + b_n}, \quad b_{n+1} = \frac{a_n + b_n}{2}. \quad (11)$$

Here he succeeds in finding the *terminatio* by an ingenious and simple idea: he determines an invariant

expression $F(a_n, b_n)$ such that

$$F(a_{n+1}, b_{n+1}) = F(a_n, b_n); \quad (12)$$

then, the *terminatio* t will satisfy the equation

$$F(a_1, b_1) = F(t, t), \quad (13)$$

which gives the value t in terms of a_1 and b_1 . For the examples (10), (11) he can state immediately the invariant expressions

$$F(a_n, b_n) = \beta a_n + \alpha b_n \quad \text{and} \quad F(a_n, b_n) = a_n \cdot b_n,$$

respectively, and he finds as the *terminatio*, using (13):

$$t = \frac{\beta a_1 + \alpha b_1}{\beta + \alpha} \quad \text{and} \quad t = \sqrt{a_1 b_1},$$

respectively.

One may remark that Gregory investigated in (6) and (11) different combinations of arithmetical, geometrical and harmonical means. One could imagine that he tried to treat other combinations of these means, but that he could not find out an algebraic expression or a geometric interpretation. In the following century the relation between the arithmetical-geometrical mean and the elliptic integrals was discovered by Lagrange, Legendre and Gauss. We know especially that Gauss studied these means in his early youth before he had any knowledge of the calculus, and that these means, later on, showed him the way to the elliptic integrals [18]. We know moreover that Pfaff, the teacher of Gauss, investigated sequences closely related to Gregory's sequence (6) [19]. Thus, we could guess that we have here an influence of Gregory's work on one of the most important theories of modern analysis, but we have no definite evidence of such connections.

The second point may be still more momentous. Gregory attempts to prove that the *terminatio* S of the polygons i_n, I_n cannot be expressed by using the traditional five "elementary" operations on i_0 and I_0 . In the preface he puts particular emphasis on this phenomenon. From his exposition we may suppose that he first had tried to "square the circle", i.e. to find such an "elementary" expression for S . But he was critical enough to recognize that the difficulties in this search could not be overcome. And realizing that the task of algebra and analysis consists as well in solving a problem as in proving, if necessary, the "impossibility" of a certain solution, he dared to try such a proof, although he did not find

any pattern for doing it. He emphasizes that since Euclid’s classification of the usual irrationalities in his tenth book, nothing of this kind has even been attempted. Of course, Leonardo Pisano had shown [20] about 1200 A.D., that a certain cubic equation cannot be solved by Euclid’s irrationalities. However, Gregory could not have had any knowledge of this investigation since it was not published before the nineteenth century. It is a testimony to Gregory’s surprising intuition that he mentions further as problems impossible in the same sense just these two: to solve the general algebraic equation and to get an n th root by solving quadratic equations.

To be sure, Gregory does not prove that it is impossible to square the circle, although this is in his mind. He approaches only a much easier problem: to prove that the area of an arbitrary circular sector S cannot be expressed in terms of the areas i_0 and I_0 by the five elementary operations — or, in modern language, that the arc tangent function as given by (8) and defined by the limiting process (6), (7), is not a combination of such algebraic functions. The foundation of his proof is the remark that two sequences (6) yield the same terminatio S whether we begin the process with i_0, I_0 or with i_1, I_1 ; therefore S depends upon i_0 and I_0 in the same way as upon i_1 and I_1 . To put it in modern language, the function satisfies the algebraic functional equation:

$$S(i_0, I_0) = S(i_1, I_1) = S\left(\sqrt{i_0 I_0}, \frac{2i_0 I_0}{i_0 + \sqrt{i_0 I_0}}\right), \tag{14}$$

i.e., $S(i_0, I_0)$ can be transformed algebraically into itself. He tries to prove that the identity (14) is impossible for any function formed only by the five elementary operations. First he removes the irrationality, introducing two suitable new variables u, v by the equations

$$i_0 = u^2(u + v), \quad I_0 = v^2(u + v).$$

Then (6) shows that

$$i_1 = uv(u + v), \quad I_1 = 2uv^2,$$

and the identity (14) becomes

$$S(u^2(u + v), v^2(u + v)) = S(uv(u + v), 2uv^2). \tag{15}$$

Now he states two properties of this identity from which he is going to deduce its impossibility for functions S of the above described algebraic type:

1) The arguments of S on the left side contain u up to the third power, while those on the right side contain u only up to the second power.

2) On the left side, both arguments are binomial, while on the right side the second one is only monomial.

Of course, Gregory is able to prove correctly by this procedure that the identity (15) cannot be satisfied by a rational integral function S of its two arguments, and even, with slightly more difficulty, that it cannot be satisfied by any rational function. However, we do not believe that the facts he offers are sufficient to furnish the proof that S is not an irrational function built up in using extraction of roots. Indeed, the algebraic factor

$$I_0 \sqrt{i_0} / \sqrt{I_0 - i_0}$$

of (8) satisfies, itself, an identity which differs from (14) only by a factor 2 in the left member, and Gregory’s considerations could be applied equally well to the modified identity. The point is that the identity (14), used as basis for his proof, implies an intrinsic difficulty: it is equivalent to the algebraic relation between $\tan \theta$ and $\tan 2\theta$ and, moreover, Gregory thinks of it only as valid in the restricted interval $0 < \theta < \frac{1}{2}\pi$.

Today, we would conclude the transcendental character of $\tan \theta$ (and, simultaneously, of the inverse function arc tangent) immediately from the periodicity of that function ($\tan \theta = \tan(\theta + \pi)$). Although such a conclusion seems to us extremely simple, it may have been difficult and remote at Gregory’s time.

A modern mathematician will highly admire Gregory’s daring attempt of “proof of impossibility” even if Gregory could not attain his aim. He will consider it a first step into a new group of mathematical questions which became extremely important in the 19th century. However, the contemporary echoes of Gregory’s undertaking were in no way favorable. First of all, Huygens criticized [21] the *Vera Quadratura* in an extremely unfavorable manner. Gregory had sent him one of the first copies. He expected his discoveries to be fully appreciated by this great mathematician who himself had done very important work on the problem of the quadrature of conics and the circle. But, unfortunately, Huygens was apparently angry that those earlier investigations were not mentioned. Thus, he put more emphasis on some claims of priority and on some objections against Gregory’s deductions than on the importance of Gregory’s new ideas and results. There is no need to report here on the unpleasant discussion which arose from this criticism [22]. We mention only the single point of importance where Huygens showed

a profounder insight. He says: even if the area of an arbitrary circular sector cannot be expressed algebraically in terms of the areas i_0, I_0 , one can still imagine such an expression to be possible for particular sectors, for example, for the whole circle itself. Gregory, obviously, had overlooked this possibility in his original publication. In his answer he tried to deduce the result for the “particular case” from that for the arbitrary sector. These endeavors could not but fail; it took more than two centuries before mathematics had developed the necessary means to prove the transcendency of π .

5 Conclusion

Surveying the importance of all these discoveries and ideas of Gregory, and realizing that the total range of his scientific work is by no means covered by our report, one may wonder why this great man did not exert more influence on the actual development of mathematics. The reason can be found in some unfortunate, almost tragical facts, in Gregory’s life which hampered his activity as well as the effectiveness of his work. After some short sojourns in London (1663 and 1668), and several years of inspiring studies in Italy (1664–1668), mostly in Padua, he was appointed Professor of Mathematics at the Scottish University of St. Andrews. At this old school, still living entirely in medieval traditions, the young scholar was rather isolated. There he was the only one who knew of the new development of mathematics. He himself abounded with new ideas, but there was no possibility to discuss or to teach them. Moreover, hardly any literature was available. Only through his correspondence with Collins whom he had met in London and who had become his close friend, could he learn what the great mathematicians in England and abroad were planning and completing.

Thus, his ideas could not find the response they deserved and he himself did not develop them as far as it might have been possible in closer contact with mathematicians of equal rank. Still worse consequences may have been involved in the lack of appreciation of his first important publication, the *Vera Quadratura*, and especially in the unkind and unjust criticism of Huygens which we have mentioned above.

Apparently, these experiences impressed the proud young Scotchman so deeply that he abandoned entirely the trend of ideas he had started so success-

fully. We can imagine that otherwise he might have applied his “convergent” pairs of sequences, as defined by recurrence formulas, to various problems and that he might have brought this important process to greater prominence in the early analysis. In fact, he afterwards used the infinite series, probably influenced by the reports he got, scantily, on Newton’s work. Yet, also here, fate did not favor him. For he was not given time and opportunity to complete and publish his investigations; and his great merits were darkened by Newton’s glory who, meanwhile, could finish his work.

Besides, Gregory had inaugurated research on differential and integral calculus without knowing what his eminent competitors were doing simultaneously in this field. He was even the first to publish, as early as 1668, a proof [23] of the “fundamental theorem,” that the two characteristic problems of the calculus, namely, to determine the slopes and to determine the areas, are inverse to one another. Also here he met misfortune; immediately afterwards there appeared Barrow’s great work *Lectiones Geometricae*, which went much farther and won all the fame. A few years later, Newton’s and Leibniz’s momentous results on the calculus became known and made obsolete the work of all their predecessors.

Gregory did not live to see this development. He had eventually taken over a professorship at the University of Edinburgh, which granted him better working opportunities. But only one year later, in the fall of 1675, he suddenly fell ill and died in his thirty-seventh year. Most of his discoveries and ideas were buried in his letters and notes or lost through his death.

References

1. Published for the Royal Society of Edinburgh, London, 1939.
2. *Pataviae*, 1667; reprinted as appendix to J. Gregory’s *Geometria pars Universalis*, Venetiae, 1668, and again in Chr. Huygens, *Opera varia*, vol. II, Lugduni Batavorum, 1724, pp. 407–462. Our report in no. 4 is based on our essay in the *Memorial* [1], pp. 468–478.
3. N. Mercator, *Logarithmotechnica*, Londini, 1668.
4. Published with a comprehensive commentary of H. W. Turnbull in the *Memorial* [1], pp. 350–359.
5. *Memorial* [1], pp. 118–122, especially p. 119 f.; cf., Turnbull’s commentary, *ibid.*, p. 124. With regard to the earlier publications of that letter, one may compare *ibid.*, pp. 25 and 29.

6. It is mentioned first, but not formulated, in Newton's letter to Oldenburg of October 24, 1676 (Newton, *Opuscula Mathematica I*, Lausannae et Genevae, 1744, pp. 328–357; see particularly p. 340) and completely stated in his *Philosophiae Naturalis Principia*, 1687, book III, lemma V, and in his *Methodus Differentialis*, 1711 (*Opuscula* [6], p. 271 ff.) — at both places for equidistant and non-equidistant ordinates.
7. See *Memorial* [1], p. 58, and Turnbull's note, p. 59.
8. In the letter quoted in [6], p. 341.
9. *Methodus Incrementorum*, Londini, 1715.
10. See *Memorial* [1], p. 131 f., and Turnbull's commentary, p. 132 f.
11. Letter of March 24, 1670, *Memorial* [1], p. 88; cf., Gregory's answer, *ibid.*, p. 92.
12. *Memorial* [1], p. 148; cf., Turnbull's commentary, p. 150 f.
13. Newton, *Opuscula I* [6], pp. 307–322, especially p. 307 f., and pp. 328–357, especially pp. 329 ff.
14. *Memorial* [1], p. 230.
15. In our essay in the *Memorial* [2], p. 469, the second of these formulas is misprinted. We may correct here some other minor misprints in that essay: p. 469, last line, read I_2 instead of $\sqrt{I_2 i_2}$; p. 471, last formula, read $b_{n+1} = \chi(a_n, b_n)$ instead of $b_{n+1} = \phi(a_n, b_n)$; p. 478, note 7, read \tanh instead of \tan on one side of the formula.
16. First part of Gregory's *Exercitationes Geometricae*, Londini, 1668.
17. See *Memorial* [1], p. 461 ff.
18. Cf., L. Schlesinger, Gauss' *Fragmente zur Theorie des arithmetisch-geometrischen Mittels*, *Nachrichten der Goettinger Gesellschaft der Wissenschaften*, 1912, and his essay in Gauss, *Werke*, vol. X, part 2, Berlin, 1933, Abhandlung 2.
19. Cf., Pfaff's letters to Gauss in Gauss, *Werke*, vol. X, part 1, Leipzig, 1917, p. 234 ff., and H. Geppert, *Mathematische Annalen*, vol. 108, 1933, p. 205 ff.
20. Published first by B. Boncompagni, *Tre scritti inediti di Leonardo Pisano*, Firenze, 1854. The proof to which we refer is reviewed comprehensively by F. Woepke, *Journal de mathématiques pures et appliquées*, vol. 19, 1854, p. 401 ff.
21. First in a review in the *Journal des Scavans*, Paris, July, 1668; cf. the references in [22].
22. One may compare, also for further literature, our essay [2] and the comprehensive report of E. J. Dijksterhuis, *Memorial* [1], pp. 478–86. The most important parts of the discussion are reprinted in Chr. Huygens, *Opera varia II* [2], pp. 463 ff., and again in the *Oeuvres complètes de Christiaan Huygens*, vol. VI, La Hage, 1895, pp. 228 ff.
23. Contained in the *Geometriae pars universalis*, Venetiae, 1668. Cf. the essay of A. Prag on this work in the *Memorial* [1], pp. 487 ff.

The Changing Concept of Change: The Derivative from Fermat to Weierstrass

JUDITH V. GRABINER

Mathematics Magazine 56 (1983), 195–206

Some years ago while teaching the history of mathematics, I asked my students to read a discussion of maxima and minima by the seventeenth-century mathematician, Pierre Fermat. To start the discussion, I asked them, “Would you please define a relative maximum?” They told me it was a place where the derivative was zero. “If that’s so,” I asked, “then what is the definition of a relative minimum?” They told me, that’s a place where the derivative is zero. “Well, in that case,” I asked, “what is the difference between a maximum and a minimum?” They replied that in the case of a maximum, the second derivative is negative.

What can we learn from this apparent victory of calculus over common sense?

I used to think that this story showed that these students did not understand the calculus, but I have come to think the opposite: they understood it very well. The students’ answers are a tribute to the power of the calculus in general, and the power of the concept of derivative in particular. Once one has been initiated into the calculus, it is hard to remember what it was like not to know what a derivative is and how to use it, and to realize that people like Fermat once had to cope with finding maxima and minima without knowing about derivatives at all.

Historically speaking, there were four steps in the development of today’s concept of the derivative, which I list here in chronological order. The derivative was first *used*; it was then *discovered*; it was then *explored and developed*; and it was finally *defined*. That is, examples of what we now recognize as derivatives first were used on an ad hoc basis in solving particular problems; then the general concept lying behind these uses was identified (as part of the

invention of the calculus); then many properties of the derivative were explained and developed in applications both to mathematics and to physics; and finally, a rigorous definition was given and the concept of derivative was embedded in a rigorous theory. I will describe the steps, and give one detailed mathematical example from each. We will then reflect on what it all means—for the teacher, for the historian, and for the mathematician.

The seventeenth-century background

Our story begins shortly after European mathematicians had become familiar once more with Greek mathematics, learned Islamic algebra, synthesized the two traditions, and struck out on their own. François Vieta invented symbolic algebra in 1591; Descartes and Fermat independently invented analytic geometry in the 1630’s. Analytic geometry meant, first, that curves could be represented by equations; conversely, it meant also that every equation determined a curve. The Greeks and Muslims had studied curves, but not that many—principally the circle and the conic sections plus a few more defined as loci. Many problems had been solved for these, including finding their tangents and areas. But since any equation could now produce a new curve, students of the geometry of curves in the early seventeenth century were suddenly confronted with an explosion of curves to consider. With these new curves, the old Greek methods of synthetic geometry were no longer sufficient. The Greeks, of course, had known how to find the tangents to circles, conic

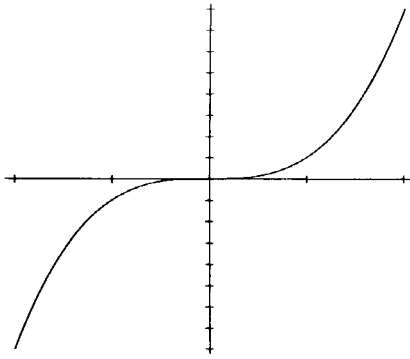


Figure 1.

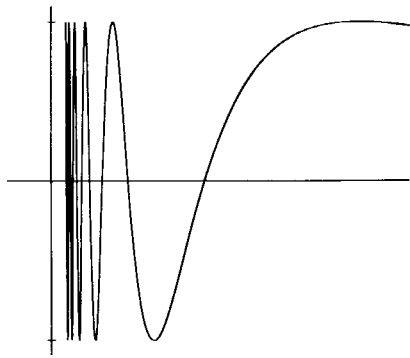


Figure 2.

sections, and some more sophisticated curves such as the spiral of Archimedes, using the methods of synthetic geometry. But how could one describe the properties of the tangent at an arbitrary point on a curve defined by a ninety-sixth degree polynomial? The Greeks had defined a tangent as a line which touches a curve without cutting it, and usually expected it to have only one point in common with the curve. How then was the tangent to be defined at the point $(0, 0)$ for a curve like $y = x^3$ (Figure 1), or to a point on a curve with many turning points (Figure 2)?

The same new curves presented new problems to the student of areas and arc lengths. The Greeks had also studied a few cases of what they called “isoperimetric” problems. For example, they asked: of all plane figures with the same perimeter, which one has the greatest area? The circle, of course, but the Greeks had no general method for solving all such problems. Seventeenth-century mathematicians hoped that the new symbolic algebra might somehow help solve all problems of maxima and minima.

Thus, though a major part of the agenda for seventeenth-century mathematicians—tangents, ar-

eas, extrema—came from the Greeks, the subject matter had been vastly extended, and the solutions would come from using the new tools: symbolic algebra and analytic geometry.

Finding maxima, minima, and tangents

We turn to the first of our four steps in the history of the derivative: its *use*, and also illustrate some of the general statements we have made. We shall look at Pierre Fermat’s method of finding maxima and minima, which dates from the 1630’s [8]. Fermat illustrated his method first in solving a simple problem, whose solution was well known: *Given a line, to divide it into two parts so that the product of the parts will be a maximum.* Let the length of the line be designated B and the first part A (Figure 3). Then the second part is $B - A$ and the product of the two parts is

$$A(B - A) = AB - A^2. \quad (1)$$

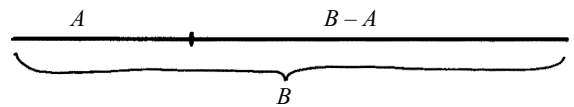


Figure 3.

Fermat had read in the writings of the Greek mathematician Pappus of Alexandria that a problem which has, in general, two solutions will have only one solution in the case of a maximum. This remark led him to his method of finding maxima and minima. Suppose in the problem just stated there is a second solution. For this solution, let the first part of the line be designated as $A + E$; the second part is then $B - (A + E) = B - A - E$. Multiplying the two parts together, we obtain for the product

$$\begin{aligned} BA + BE - A^2 - AE - EA - E^2 \\ = AB - A^2 - 2AE + BE - E^2. \end{aligned} \quad (2)$$

Following Pappus’ principle for the maximum, instead of two solutions, there is only one. So we set the two products (1) and (2) “sort of” equal; that is, we formulate what Fermat called the pseudo-equality:

$$AB - A^2 = AB - A^2 - 2AE + BE - E^2.$$

Simplifying, we obtain $2AE + E^2 = BE$ and $2A + E = B$. Now Fermat said, with no justification and

no ceremony, “suppress E .” Thus he obtained $A = B/2$, which indeed gives the maximum sought. He concluded, “We can hardly expect a more general method.” And, of course, he was right.

Notice that Fermat did not call E infinitely small, or vanishing, or a limit; he did not explain why he could first divide by E (treating it as non-zero) and then throw it out (treating it as zero). Furthermore, he did not explain what he was doing as a special case of a more general concept, be it derivative, rate of change, or even slope of tangent. He did not even understand the relationship between his maximum-minimum method and the way one found tangents; in fact he followed his treatment of maxima and minima by saying that the same method—that is, adding E , doing the algebra, then suppressing E —could be used to find tangents [8, p. 223].

Though the considerations that led Fermat to his method may seem surprising to us, he did devise a method of finding extrema that worked, and it gave results that were far from trivial. For instance, Fermat applied his method to optics. Assuming that a ray of light which goes from one medium to another always takes the quickest path (what we now call the Fermat least-time principle), he used his method to compute the path taking minimal time. Thus he showed that his least-time principle yields Snell’s law of refraction [7] [12, pp. 387–390].

Though Fermat did not publish his method of maxima and minima, it became well known through correspondence and was widely used. After mathematicians had become familiar with a variety of examples, a pattern emerged from the solutions by Fermat’s method to maximum-minimum problems. In 1659, Johann Hudde gave a general verbal formulation of this pattern [3, p. 186], which, in modern notation, states that, *given a polynomial of the form*

$$y = \sum_{k=0}^n a_k x^k$$

there is a maximum or minimum when

$$\sum_{k=1}^n k a_k x^{k-1} = 0.$$

Of even greater interest than the problem of extrema in the seventeenth century was the finding of tangents. Here the tangent was usually thought of as a secant for which the two points came closer and closer together until they coincided. Precisely what it meant for a secant to “become” a tangent was never

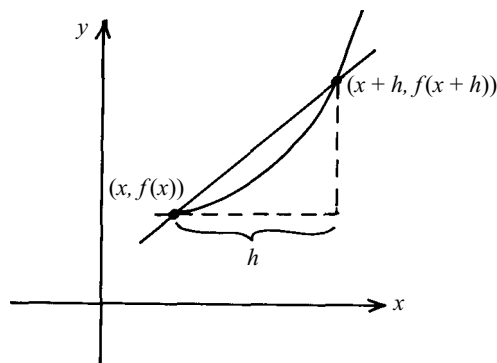


Figure 4.

completely explained. Nevertheless, methods based on this approach worked. Given the equation of a curve $y = f(x)$, Fermat, Descartes, John Wallis, Isaac Barrow, and many other seventeenth-century mathematicians were able to find the tangent. The method involves considering, and computing, the slope of the secant,

$$\frac{f(x+h) - f(x)}{h},$$

doing the algebra required by the formula for $f(x+h)$ in the numerator, then dividing by h . The diagram in Figure 4 then suggests that when the quantity h vanishes, the secant becomes the tangent, so that neglecting h in the expression for the slope of the secant gives the slope of the tangent. Again, a general pattern for the equations of slopes of tangents soon became apparent, and a rule analogous to Hudde’s rule for maxima and minima was stated by several people, including René Sluse, Hudde, and Christiaan Huygens [3, pp. 185–186].

By the year 1660, both the computational and the geometric relationships between the problem of extrema and the problem of tangents were clearly understood; that is, a maximum was found by computing the slope of the tangent, according to the rule, and asking when it was zero. While in 1660 there was not yet a general concept of derivative, there was a general method for solving one type of geometric problem. However, the relationship of the tangent to other geometric concepts—area, for instance—was not understood, and there was no completely satisfactory definition of tangent. Nevertheless, there was a wealth of methods for solving problems that we now solve by using the calculus, and in retrospect, it would seem to be possible to generalize those methods. Thus in this context it is natural to ask, how did the derivative as we know it come to be?

It is sometimes said that the idea of the derivative was motivated chiefly by physics. Newton, after all, invented both the calculus and a great deal of the physics of motion. Indeed, already in the Middle Ages, physicists, following Aristotle who had made “change” the central concept in his physics, logically analyzed and classified the different ways a variable could change. In particular, something could change uniformly or non-uniformly; if non-uniformly, it could change uniformly-non-uniformly or non-uniformly-non-uniformly, etc. [3, pp. 73–74]. These medieval classifications of variation helped to lead Galileo in 1638, without benefit of calculus, to his successful treatment of uniformly accelerated motion. Motion, then, could be studied scientifically. Were such studies the origin and purpose of the calculus? The answer is no. However plausible this suggestion may sound, and however important physics was in the later development of the calculus, physical questions were in fact neither the immediate motivation nor the first application of the calculus. Certainly they prepared people’s thoughts for some of the properties of the derivative, and for the introduction into mathematics of the concept of change. But the immediate motivation for the general concept of derivative—as opposed to specific examples like speed or slope of tangent—did not come from physics. The first problems to be solved, as well as the first applications, occurred in mathematics, especially geometry (see [1, chapter 7]; see also [3; chapters 4–5], and, for Newton, [17]). The concept of derivative then developed gradually, together with the ideas of extrema, tangent, area, limit, continuity, and function, and it interacted with these ideas in some unexpected ways.

Tangents, areas, and rates of change

In the latter third of the seventeenth century, Newton and Leibniz, each independently, invented the calculus. By “inventing the calculus” I mean that they did three things. First, they took the wealth of methods that already existed for finding tangents, extrema, and areas, and they subsumed all these methods under the heading of two general concepts, the concepts which we now call *derivative* and *integral*. Second, Newton and Leibniz each worked out a notation which made it easy, almost automatic, to use these general concepts. (We still use Newton’s \dot{x} and we still use Leibniz’s dy/dx and $\int ydx$.) Third, Newton

and Leibniz each gave an argument to prove what we now call the Fundamental Theorem of Calculus: the derivative and the integral are mutually inverse. Newton called our “derivative” a *fluxion*—a rate of flux or change; Leibniz saw the derivative as a ratio of infinitesimal differences and called it the *differential quotient*. But whatever terms were used, the concept of derivative was now embedded in a general subject—the calculus—and its relationship to the other basic concept, which Leibniz called the integral, was now understood. Thus we have reached the stage I have called *discovery*.

Let us look at an early Newtonian version of the Fundamental Theorem [13, sections 54-5, p. 23]. This will illustrate how Newton presented the calculus in 1669, and also illustrate both the strengths and weaknesses of the understanding of the derivative in this period.

Consider with Newton a curve under which the area up to the point $D = (x, y)$ is given by z (see Figure 5). His argument is general: “Assume any relation betwixt x and z that you please,” he then proceeded to find y . The example he used is

$$z = \frac{n}{m+n}ax^{(m+n)/n};$$

however, it will be sufficient to use $z = x^3$ to illustrate his argument.

In the diagram in Figure 5, the auxiliary line bd is chosen so that $Bb = o$, where o is not zero. Newton then specified that $BK = v$ should be chosen so that area $BbHK = \text{area } BbdD$. Thus $ov = \text{area } BbdD$. Now, as x increases to $x + o$, the change in the area z is given by

$$\begin{aligned} z(x+o) - z(x) &= x^3 + 3x^2o + 3xo^2 + o^3 - x^3 \\ &= 3x^2o + 3xo^2 + o^3, \end{aligned}$$

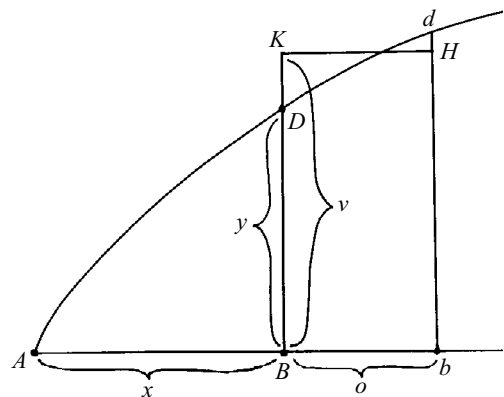


Figure 5.

which, by the definition of v , is equal to ov . Now since

$$3x^2o + 3xo^2 + o^3 = ov,$$

dividing by o produces $3x^2 + 3ox + o^2 = v$. Now, said Newton, “If we suppose Bb to be diminished infinitely and to vanish, or o to be nothing, v and y in that case will be equal and the terms which are multiplied by o will vanish: so that there will remain . . .” $3x^2 = y$.

What has he shown? Since

$$\frac{z(x+o) - z(x)}{o}$$

is the rate at which the area z changes, that rate is given by the ordinate y . Moreover, we recognize that $3x^2$ would be the slope of the tangent to the curve $z = x^3$. Newton went on to say that the argument can be reversed; thus the converse holds too. We see that derivatives are fundamentally involved in areas as well as tangents, so the concept of derivative helps us to see that these two problems are mutually inverse. Leibniz gave analogous arguments on this same point (see, e.g. [16, pp. 282-284]).

Newton and Leibniz did not, of course, have the last word on the concept of derivative. Though each man had the most useful properties of the concept, there were still many unanswered questions. In particular, what, exactly, is a differential quotient? Some disciples of Leibniz, notably Johann Bernoulli and his pupil the Marquis de l’Hospital, said a differential quotient was a ratio of infinitesimals; after all, that is the way it was calculated. But infinitesimals, as seventeenth-century mathematicians were well aware, do not obey the Archimedean axiom. Since the Archimedean axiom was the basis for the Greek theory of ratios, which was, in turn, the basis of arithmetic, algebra, and geometry for seventeenth-century mathematicians, non-Archimedean objects were viewed with some suspicion. Again, what is a fluxion? Though it can be understood intuitively as a velocity, the proofs Newton gave in his 1671 *Method of Fluxions* all involved an “indefinitely small quantity o ”, [14, pp. 32-33] which raises many of the same problems that the o which “vanishes” raised in the Newtonian example of 1669 we saw above. In particular, what is the status of that little o ? Is it zero? If so, how can we divide by it? If it is not zero, aren’t we making an error when we throw it away? These questions had already been posed in Newton’s and Leibniz’s time. To avoid such problems, Newton said in 1687 that quantities defined in the way that $3x^2$ was defined in our example were the *limit* of

the ratio of vanishing increments. This sounds good, but Newton’s understanding of the term “limit” was not ours. Newton in his *Principia* (1687) described limits as “ultimate ratios”—that is, the value of the ratio of those vanishing quantities just when they are vanishing. He said, “Those ultimate ratios with which quantities vanish are not truly the ratios of ultimate quantities, but limits towards which the ratios of quantities decreasing without limit do always converge; and to which they approach nearer than by any given difference, but never go beyond, nor in effect attain to, till the quantities are diminished in infinitum” [15, Book I, Scholium to Lemma XI, p. 39].

Notice the phrase “but never go beyond”—so a variable cannot oscillate about its limit. By “limit” Newton seems to have had in mind “bound”, and mathematicians of his time often cite the particular example of the circle as the limit of inscribed polygons. Also, Newton said, “nor . . . attain to, till the quantities are diminished in infinitum.” This raises a central issue: it was often asked whether a variable quantity ever actually reached its limit. If it did not, wasn’t there an error? Newton did not help clarify this when he stated as a theorem that “Quantities and the ratios of quantities which in any finite time converge continually to equality, and before the end of that time approach nearer to each other than by any given difference, become ultimately equal” [15, Book I, Lemma I, p. 29]. What does “become ultimately equal” mean? It was not really clear in the eighteenth century, let alone the seventeenth.

In 1734, George Berkeley, Bishop of Cloyne, attacked the calculus on precisely this point. Scientists, he said, attack religion for being unreasonable; well, let them improve their own reasoning first. A quantity is either zero or not; there is nothing in between. And Berkeley characterized the mathematicians of his time as men “rather accustomed to compute, than to think” [2].

Perhaps Berkeley was right, but most mathematicians were not greatly concerned. The concepts of differential quotient and integral, concepts made more effective by Leibniz’s notation and by the Fundamental Theorem, had enormous power. For eighteenth-century mathematicians, especially those on the Continent where the greatest achievements occurred, it was enough that the concepts of the calculus were understood sufficiently well to be applied to solve a large number of problems, both in mathematics and in physics. So, we come to our third stage: *exploration and development*.

Differential equations, Taylor series, and functions

Newton had stated his three laws of motion in words, and derived his physics from those laws by means of synthetic geometry [15]. Newton's second law stated: "*The change of motion [our 'momentum'] is proportional to the motive force impressed, and is made in the direction of the [straight] line in which that force is impressed*" [15, p. 13]. Once translated into the language of the calculus, this law provided physicists with an instrument of physical discovery of tremendous power—because of the power of the concept of the derivative.

To illustrate, if F is force and x distance (so $m\dot{x}$ is momentum and, for constant mass, $m\ddot{x}$ the rate of change of momentum), then Newton's second law takes the form $F = m\ddot{x}$. Hooke's law of elasticity (when an elastic body is distorted the restoring force is proportional to the distance [in the opposite direction] of the distortion) takes the algebraic form $F = -kx$. By equating these expressions for force, Euler in 1739 could easily both state and solve the differential equation $m\ddot{x} + kx = 0$ which describes the motion of a vibrating spring [10, p. 482]. It was mathematically surprising, and physically interesting, that the solution to that differential equation involves sines and cosines.

An analogous, but considerably more sophisticated problem, was the statement and solution of the partial differential equation for the vibrating string. In modern notation, this is

$$\frac{\partial^2 y}{\partial t^2} = \frac{T}{\mu} \frac{\partial^2 y}{\partial x^2},$$

where T is the tension in the string and μ is its mass per unit length. The question of how the solutions to this partial differential equation behaved was investigated by such men as d'Alembert, Daniel Bernoulli, and Leonhard Euler, and led to extensive discussions about the nature of continuity, and to an expansion of the notion of function from formulas to more general dependence relations [10, pp. 502–514], [16, pp. 367–368]. Discussions surrounding the problem of the vibrating string illustrate the unexpected ways that discoveries in mathematics and physics can interact ([16, pp. 351–368] has good selections from the original papers). Numerous other examples could be cited, from the use of infinite-series approximations in celestial mechanics to the dynamics of rigid bodies, to show that by the mid-eighteenth century the differential equation had become the most useful

mathematical tool in the history of physics.

Another useful tool was the Taylor series, developed in part to help solve differential equations. In 1715, Brook Taylor, arguing from the properties of finite differences, wrote an equation expressing what we would write as $f(x+h)$ in terms of $f(x)$ and its quotients of differences of various orders. He then let the differences get small, passed to the limit, and gave the formula that still bears his name: the Taylor series. (Actually, James Gregory and Newton had anticipated this discovery, but Taylor's work was more directly influential.) The importance of this property of derivatives was soon recognized, notably by Colin Maclaurin (who has a special case of it named after him), by Euler, and by Joseph-Louis Lagrange. In their hands, the Taylor series became a powerful tool in studying functions and in approximating the solution of equations.

But beyond this, the study of Taylor series provided new insights into the nature of the derivative. In 1755, Euler, in his study of power series, had said that for any power series,

$$a + bx + cx^2 + dx^3 + \dots,$$

one could find x sufficiently small so that if one broke off the series after some particular term—say x^2 —the x^2 term would exceed, in absolute value, the sum of the entire remainder of the series [6, section 122]. Though Euler did not prove this—he must have thought it obvious since he usually worked with series with finite coefficients—he applied it to great advantage. For instance, he could use it to analyze the nature of maxima and minima. Consider, for definiteness, the case of maxima. If $f(x)$ is a relative maximum, then by definition, for small h ,

$$f(x-h) < f(x) \text{ and } f(x+h) < f(x).$$

Taylor's theorem gives, for these inequalities,

$$\begin{aligned} f(x-h) &= f(x) - h \frac{df(x)}{dx} + h^2 \frac{d^2 f(x)}{dx^2} - \dots \\ &< f(x); \end{aligned} \quad (3)$$

$$\begin{aligned} f(x+h) &= f(x) + h \frac{df(x)}{dx} + h^2 \frac{d^2 f(x)}{dx^2} + \dots \\ &< f(x). \end{aligned} \quad (4)$$

Now if h is so small that $hdf(x)/dx$ dominates the rest of the terms, the only way that both of the inequalities (3) and (4) can be satisfied is for $df(x)/dx$

to be zero. Thus the differential quotient is zero for a relative maximum. Furthermore, Euler argued, since h^2 is always positive, if $d^2f(x)/dx^2 \neq 0$, the only way both inequalities can be satisfied is for $d^2f(x)/dx^2$ to be negative. This is because the h^2 term dominates the rest of the series—unless $d^2f(x)/dx^2$ is itself zero, in which case we must go on and think about even higher-order differential quotients. This analysis, first given and demonstrated geometrically by Maclaurin, was worked out in full analytic detail by Euler [6, sections 253–254], [9, pp. 117–118]. It is typical of Euler’s ability to choose computations that produce insight into fundamental concepts. It assumes, of course, that the function in question has a Taylor series, an assumption which Euler made without proof for many functions; it assumes also that the function is uniquely the sum of its Taylor series, which Euler took for granted. Nevertheless, this analysis is a beautiful example of the exploration and development of the concept of the differential quotient of first, second, and n th orders—a development which completely solves the problem of characterizing maxima and minima, a problem which goes back to the Greeks.

Lagrange and the derivative as a function

Though Euler did a good job analyzing maxima and minima, he brought little further understanding of the nature of the differential quotient. The new importance given to Taylor series meant that one had to be concerned not only about first and second differential quotients, but about differential quotients of any order.

The first person to take these questions seriously was Lagrange. In the 1770’s, Lagrange was impressed with what Euler had been able to achieve by Taylor-series manipulations with differential quotients, but Lagrange soon became concerned about the logical inadequacy of all the existing justifications for the calculus. In particular, Lagrange wrote in 1797 that the Newtonian limit-concept was not clear enough to be the foundation for a branch of mathematics. Moreover, in not allowing variables to surpass their limits, Lagrange thought the limit-concept too restrictive. Instead, he said, the calculus should be reduced to algebra, a subject whose foundations in the eighteenth century were generally thought to be sound [11, pp. 15–16].

The algebra Lagrange had in mind was what he called the algebra of infinite series, because Lagrange was convinced that infinite series were part of algebra. Just as arithmetic deals with infinite decimal fractions without ceasing to be arithmetic, Lagrange thought, so algebra deals with infinite algebraic expressions without ceasing to be algebra. Lagrange believed that expanding $f(x + h)$ into a power series in h was always an algebraic process. It is obviously algebraic when one turns $1/(1 - x)$ into a power series by dividing. And Euler had found, by manipulating formulas, infinite power-series expansions for functions like $\sin x$, $\cos x$, e^x . If functions like those have power-series expansions, perhaps everything could be reduced to algebra. Euler, in his book *Introduction to the analysis of the infinite (Introductio in analysin infinitorum, 1748)*, had studied infinite series, infinite products, and infinite continued fractions by what he thought of as purely algebraic methods. For instance, he converted infinite series into infinite products by treating a series as a very long polynomial. Euler thought that this work was purely algebraic, and—what is crucial here—Lagrange also thought Euler’s methods were purely algebraic. So Lagrange tried to make the calculus rigorous by reducing it to the algebra of infinite series.

Lagrange stated in 1797, and thought he had proved, that any function (that is, any analytic expression, finite or infinite) had a power-series expansion:

$$f(x + h) = f(x) + p(x)h + q(x)h^2 + r(x)h^3 + \dots, \quad (5)$$

except, possibly, for a finite number of isolated values of x . He then defined a new function, the coefficient of the linear term in h (which is $p(x)$ in the expansion shown in (5)) and called it the first *derived function* of $f(x)$. Lagrange’s term “derived function” (*fonction dérivée*) is the origin of our term “derivative.” Lagrange introduced a new notation, $f'(x)$, for that function. He defined $f''(x)$ to be the first derived function of $f'(x)$, and so on, recursively. Finally, using these definitions, he proved that, in the expansion (5) above,

$$q(x) = f''(x)/2, \quad r(x) = f'''(x)/6,$$

and so on [11, chapter 2].

What was new about Lagrange’s definition? The concept of *function*—whether simply an algebraic expression (possibly infinite) or, more generally,

any dependence relation—helps free the concept of derivative from the earlier ill-defined notions. Newton’s explanation of a fluxion as a rate of change appeared to involve the concept of motion in mathematics; moreover, a fluxion seemed to be a different kind of object than the flowing quantity whose fluxion it was. For Leibniz, the differential quotient had been the quotient of vanishingly small differences; the second differential quotient, of even smaller differences. Bishop Berkeley, in his attack on the calculus, had made fun of these earlier concepts, calling vanishing increments “ghosts of departed quantities” [2, section 35]. But since, for Lagrange, the derivative was a function, it was now the same sort of object as the original function. The second derivative is precisely the same sort of object as the first derivative; even the n th derivative is simply another function, defined as the coefficient of h in the Taylor series for $f^{(n-1)}(x+h)$. Lagrange’s notation $f'(x)$ was designed precisely to make this point.

We cannot fully accept Lagrange’s definition of the derivative, since it assumes that every differentiable function is the sum of a Taylor series and thus has infinitely many derivatives. Nevertheless, that definition led Lagrange to a number of important properties of the derivative. He used his definition together with Euler’s criterion for using truncated power series in approximations to give a most useful characterization of the derivative of a function [9, p. 116, pp. 118–121]:

$$f(x+h) = f(x) + hf'(x) + hH,$$

where H goes to zero with h . (I call this the *Lagrange property of the derivative*.) Lagrange interpreted the phrase “ H goes to zero with h ” in terms of inequalities. That is, he wrote that,

Given D , h can be chosen so that $f(x+h) - f(x)$ lies between $h(f'(x) - D)$ and $h(f'(x) + D)$. (6)

Formula (6) is recognizably close to the modern delta-epsilon definition of the derivative.

Lagrange used inequality (6) to prove theorems. For instance, he proved that a function with positive derivative on an interval is increasing there, and used that theorem to derive the Lagrange remainder of the Taylor series [9, pp. 122–127], [11, pp. 78–85]. Furthermore, he said, considerations like inequality (6) are what make possible applications of the differential calculus to a whole range of problems in mechanics, in geometry, and, as we have described, the problem of maxima and minima (which Lagrange solved using the Taylor series remainder which bears his name [11, pp. 233–237]).

In Lagrange’s 1797 work, then, the derivative is defined by its position in the Taylor series—a strange definition to us. But the derivative is also *described* as satisfying what we recognize as the appropriate delta-epsilon inequality, and Lagrange applied this inequality and its n th-order analogue, the Lagrange remainder, to solve problems about tangents, orders of contact between curves, and extrema. Here the derivative was clearly a function, rather than a ratio or a speed.

Still, it is a lot to assume that a function has a Taylor series if one wants to define only one derivative. Further, Lagrange was wrong about the algebra of infinite series. As Cauchy pointed out in 1821, the algebra of finite quantities cannot automatically be extended to infinite processes. And, as Cauchy also pointed out, manipulating Taylor series is not foolproof. For instance, e^{-1/x^2} has a zero Taylor series about $x = 0$, but the function is not identically zero. For these reasons, Cauchy rejected Lagrange’s definition of derivative and substituted his own.

Definitions, rigor, and proofs

Now we come to the last stage in our chronological list: *definition*. In 1823, Cauchy defined the derivative of $f(x)$ as the limit, when it exists, of the quotient of differences $(f(x+h) - f(x))/h$ as h goes to zero [4, pp. 22–23]. But Cauchy understood “limit” differently than had his predecessors. Cauchy entirely avoided the question of whether a variable ever reached its limit; he just didn’t discuss it. Also, knowing an absolute value when he saw one, Cauchy followed Simon l’Huilier and S.-F. Lacroix in abandoning the restriction that variables never surpass their limits. Finally, though Cauchy, like Newton and d’Alembert before him, gave his definition of limit in words, Cauchy’s understanding of limit (most of the time, at least) was algebraic. By this, I mean that when Cauchy needed a limit property in a proof, he used the algebraic inequality characterization of limit. Cauchy’s proof of the mean value theorem for derivatives illustrates this. First he proved a theorem which states: *if $f(x)$ is continuous on $[x, x+a]$, then*

$$\min_{[x, x+a]} f'(x) \leq \frac{f(x+a) - f(x)}{a} \leq \max_{[x, x+a]} f'(x).$$

The first step in his proof is [4, p. 44]:

Let δ, ϵ be two very small numbers; the first is chosen so that for all [absolute] values of h less than δ

and for any value of x [on the given interval], the ratio $(f(x+h) - f(x))/h$ will always be greater than $f'(x) - \epsilon$ and less than $f'(x) + \epsilon$.

(The notation in this quote is Cauchy's, except that I have substituted h for the i he used for the increment.) Assuming the intermediate-value theorem for continuous functions, which Cauchy had proved in 1821, the mean-value theorem is an easy corollary of (7) [4, pp. 44–45], [9, pp. 168–170].

Cauchy took the inequality-characterization of the derivative from Lagrange (possibly via an 1806 paper of A.-M. Ampere [9, pp. 127–132]). But Cauchy made that characterization into a definition of derivative. Cauchy also took from Lagrange the name derivative and the notation $f'(x)$, emphasizing the functional nature of the derivative. And, as I have shown in detail elsewhere [9, chapter 5], Cauchy adapted and improved Lagrange's inequality proof-methods to prove results like the mean-value theorem, proof-methods now justified by Cauchy's definition of derivative.

But of course, with the new and more rigorous definition, Cauchy went far beyond Lagrange. For instance, using his concept of limit to define the integral as the limit of sums, Cauchy made a good first approximation to a real proof of the Fundamental Theorem of Calculus [9, pp. 171–175], [4, pp. 122–125, 151–152]. And it was Cauchy who not only raised the question, but gave the first proof, of the existence of a solution to a differential equation [9, pp. 158–159].

After Cauchy, the calculus itself was viewed differently. It was seen as a rigorous subject, with good definitions and with theorems whose proofs were based on those definitions, rather than merely as a set of powerful methods. Not only did Cauchy's new rigor establish the earlier results on a firm foundation, but it also provided a framework for a wealth of new results, some of which could not even be formulated before Cauchy's work.

Of course, Cauchy did not himself solve all the problems occasioned by his work. In particular, Cauchy's definition of the derivative suffers from one deficiency of which he was unaware. Given an ϵ , he chose a δ which he assumed would work for any x . That is, he assumed that the quotient of differences converged uniformly to its limit. It was not until the 1840's that G. G. Stokes, V. Seidel, K. Weierstrass, and Cauchy himself worked out the distinction between convergence and uniform convergence. After all, in order to make this distinction, one first needs a clear and algebraic understanding of what

a limit is—the understanding Cauchy himself had provided.

In the 1850's, Karl Weierstrass began to lecture at the University of Berlin. In his lectures, Weierstrass made algebraic inequalities replace words in theorems in analysis, and used his own clear distinction between pointwise and uniform convergence along with Cauchy's delta-epsilon techniques to present a systematic and thoroughly rigorous treatment of the calculus. Though Weierstrass did not publish his lectures, his students—H. A. Schwartz, G. Mittag-Leffler, E. Heine, S. Pincherlé, Sonya Kowalevsky, Georg Cantor, to name a few—disseminated Weierstrassian rigor to the mathematical centers of Europe. Thus although our modern delta-epsilon definition of derivative cannot be quoted from the works of Weierstrass, it is in fact the work of Weierstrass [3, pp. 284–287]. The rigorous understanding brought to the concept of the derivative by Weierstrass is signaled by his publication in 1872 of an example of an everywhere continuous, nowhere differentiable function. This is a far cry from merely acknowledging that derivatives might not always exist, and the example shows a complete mastery of the concepts of derivative, limit, and existence of limit [3, p. 285].

Historical development versus textbook exposition

The span of time from Fermat to Weierstrass is over two hundred years. How did the concept of derivative develop? Fermat implicitly used it; Newton and Leibniz discovered it; Taylor, Euler, Maclaurin developed it; Lagrange named and characterized it; and only at the end of this long period of development did Cauchy and Weierstrass define it. This is certainly a complete reversal of the usual order of textbook exposition in mathematics, where one starts with a definition, then explores some results, and only then suggests applications.

This point is important for the teacher of mathematics: the historical order of development of the derivative is the reverse of the usual order of textbook exposition. Knowing the history helps us as we teach about derivatives. We should put ourselves where mathematicians were before Fermat, and where our beginning students are now—back on the other side, before we had any concept of derivative, and also before we knew the many uses of derivatives. Seeing the historical origins of a concept helps motivate the concept, which we—along

with Newton and Leibniz—want for the problems it helps to solve. Knowing the historical order also helps to motivate the rigorous definition—which we, like Cauchy and Weierstrass, want in order to justify the uses of the derivative, and to show precisely when derivatives exist and when they do not. We need to remember that the rigorous definition is often the end, rather than the beginning, of a subject.

The real historical development of mathematics—the order of discovery—reveals the creative mathematician at work, and it is creation that makes doing mathematics so exciting. The order of exposition, on the other hand, is what gives mathematics its characteristic logical structure and its incomparable deductive certainty. Unfortunately, once the classic exposition has been given, the order of discovery is often forgotten. The task of the historian is to recapture the order of discovery: not as we think it might have been, not as we think it should have been, but as it really was. And this is the purpose of the story we have just told of the derivative from Fermat to Weierstrass.

References

- Margaret Baron, *Origins of the Infinitesimal Calculus*, Pergamon, Oxford, 1969.
- George Berkeley, *The Analyst, or a Discourse Addressed to an Infidel Mathematician*, 1734, in A. A. Luce and T. R. Jessop, eds., *The Works of George Berkeley*, Nelson, London, 1951 (some excerpts appear in [16, pp. 333–338]).
- Carl Boyer, *History of the Calculus and Its Conceptual Development*, Dover, New York, 1959.
- A.-L. Cauchy, *Résumé des leçons données à l'école royale polytechnique sur le calcul infinitésimal*, Paris, 1823, in *Oeuvres complètes d'Augustin Cauchy*, Gauthier-Villars, Paris, 1882–83, series 2, vol. 4.
- Pierre Dugac, *Fondements d'analyse*, in J. Dieudonné, *Abrégé d'histoire des mathématiques*, 1700–1900, 2 vols., Hermann, Paris, 1978.
- Leonhard Euler, *Institutiones calculi differentialis*, St. Petersburg, 1755, in *Opera omnia*, Teubner, Leipzig, Berlin, and Zurich, 1911–, series 1, vol. 10.
- Pierre Fermat, *Analysis ad refractiones*, 1661. In *Oeuvres de Fermat*, ed., C. Henry and P. Tannery, 4 vols., Paris, 1891–1912; *Supplement*, ed. C. de Waard, Paris, 1922, vol. 1, pp. 170–172.
- , *Methodus ad disquirendam maximam et minimam et de tangentibus linearum curvarum*, *Oeuvres*, vol. 1, pp. 133–136. Excerpted in English in [16, pp. 222–225].
- Judith V. Grabiner, *The Origins of Cauchy's Rigorous Calculus*, MIT Press, Cambridge and London, 1981.
- Morris Kline, *Mathematical Thought from Ancient to Modern Times*, Oxford, New York, 1972.
- J.-L. Lagrange, *Théorie des fonctions analytiques*, Paris, 2nd edition, 1813, in *Oeuvres de Lagrange*, ed. M. Serret, Gauthier-Villars, Paris, 1867–1892, vol. 9.
- Michael S. Mahoney, *The Mathematical Career of Pierre de Fermat, 1601–1665*, Princeton University Press, Princeton, 1973.
- Isaac Newton, *Of Analysis by Equations of an Infimite Number of Terms* [1669], in D. T. Whiteside, ed., *Mathematical Works of Isaac Newton*, Johnson, New York and London, 1964, vol. 1, pp. 3–25.
- , *Method of Fluxions* [1671], in D. T. Whiteside, ed., *Mathematical Works of Isaac Newton*, vol. 1, pp. 29–139.
- , *Mathematical Principles of Natural Philosophy*, tr. A. Motte, ed. F. Cajori, University of California Press, Berkeley, 1934.
- D. J. Struik, *Source Book in Mathematics, 1200–1800*, Harvard University Press, Cambridge, MA, 1969.
- D. T. Whiteside, ed, *The Mathematical Papers of Isaac Newton*, Cambridge University Press, 1967–1982.

The Crooked Made Straight: Roberval and Newton on Tangents

PAUL R. WOLFSON

American Mathematical Monthly 108 (2001), 206–216

1 Introduction

In October 1665, about two years after he had first read a mathematics book, Isaac Newton began investigating a method for finding the tangents to “mechanical” curves. He can have known only vaguely that he was following a path trod previously by several outstanding mathematicians, Torricelli, Descartes, Roberval, and Barrow among them. In his ignorance of the details of their work, Newton stumbled before setting himself firmly on the way to his calculus. As he progressed, he overcame the inadequate mathematical language that had kept others from expressing—sometimes from even thinking—their ideas clearly.

Newton’s method found tangents by regarding a curve as the trajectory of a moving particle, so that the velocity vector lies along the tangent. Sometimes one can easily find the velocity vector, however, by decomposing the given motion into simpler ones with known velocity vectors. This method of finding tangents to curves by decomposing the velocity vector is often called *the kinematic method*. Newton’s first manuscript on the kinematic method included three examples of curves that had traditionally been described by the composition of motions: the spiral of Archimedes, the cycloid, and the quadratrix. In addition to these mechanical curves, described as trajectories, Newton also discussed the ellipse, a so-called geometrical curve.

Newton had not been the first to consider composition of motions in general or any of these particular examples. Of course, the general idea of composition of motions goes back to the ancient Greeks, as the examples of the Archimedean spiral and the

quadratrix show. (The epicyclic paths of Ptolemaic astronomy give other ancient examples of curves generated by two or more motions.) In his *Two New Sciences* (1638), Galileo decomposed a projectile’s motion into a uniform horizontal motion and a uniformly accelerated vertical motion in order to show that the trajectory is a parabola. In his *On the Motion of Heavy Bodies* (1644), Torricelli reversed the line of argument: by treating the parabola as a trajectory formed from the composition of two motions, he explained how to determine its tangent at any point. Although the parabola was of special interest to him, Torricelli also discussed the Archimedean spiral and the cycloid in similar terms.

That same year, Descartes published his *Principles of Philosophy*. There, he discussed the decomposition of movements in a general way and illustrated the idea with a description of the cycloid. Descartes went beyond generalities, however: in a fragment published posthumously, he actually used the kinematic method to find the tangent to the quadratrix. By this time, too, Roberval had lectured on the kinematic method and had communicated it to Fermat. He took an insight shared by a few leading mathematicians and made it into a successful method.

2 What Roberval knew

Giles Personne de Roberval (1602–1675) expounded his method at the Collège Royal in Paris, where he held the Ramus chair from 1643 until his death. One of his students was François de Verdus, whose notes formed the basis for Roberval’s presentation to the Académie Royale des Sciences. The Académie pub-

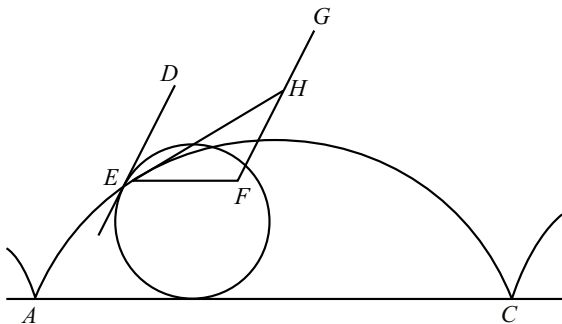


Figure 1.

lished it in 1693 as *Observations on the composition of Movements and on the means of finding Tangents to curved lines* [14]. The work begins with a discussion of the nature of composite motions. Roberval observed that uniform rectilinear motions combine by the parallelogram law (vector addition). To find tangents by decomposing motions, Roberval offered the following

General Rule. By the specific properties of the given curve, examine the divers motions of a point in the place where you want the tangent, and compose all the motions to a single one; draw the line of direction of motion, and you have the tangent. [14, p. 24]

To see what Roberval meant, we may begin with an example. Like Torricelli and others, Roberval discussed the cycloid (Figure 1). He defined it in the usual manner as the path traced by a point E on a circle that rolls on a straight line. It follows, he observed, that the instantaneous motion of E can be decomposed into a linear motion parallel to the base line and an instantaneous circular motion given by the tangent to the circle. Therefore, to draw the tangent to the cycloid at E , he used essentially the following steps: draw $EF \parallel AC$ and draw FG parallel to the tangent ED to the circle. On FG take H so that

$$AC : \text{circumference of the circle} :: EF : FH.$$

Then EH is the tangent.

This example, considered again by Barrow and Newton, does not yet make clear the subtlety of Roberval’s method. Consider, therefore, another of Roberval’s examples, the ellipse (Figure 2). From its definition as the locus of points F for which

$$FB + FA = \text{constant}$$

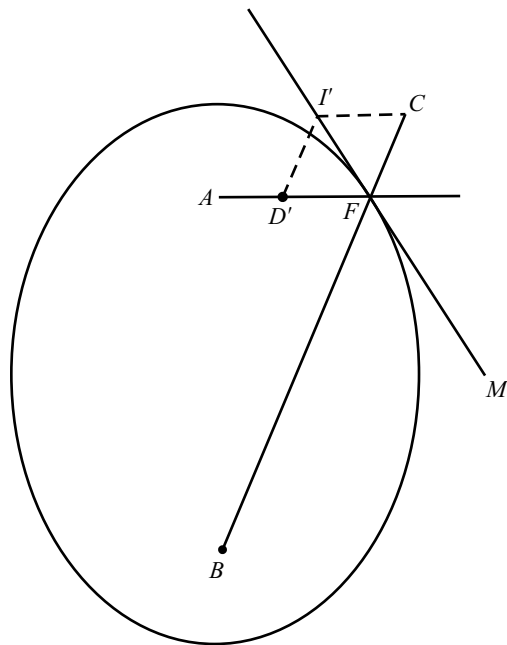


Figure 2.

it follows, he observed, that F moves so that either it recedes from B in FC and approaches A in FA , or *vice versa*. Since F recedes from one point as much as it approaches the other, the motion of F can be decomposed into two motions, represented by lengths FC and FD' , for example, and the composite motion is represented by the diagonal of the parallelogram $FCI'D'$. Therefore, the line $I'FM$ that bisects $\angle AFC$ is the tangent line.

Here, as in the example of the cycloid, Roberval was apparently recombining component velocities by vector addition. More specific than Roberval’s General Rule, this method has often been called the Parallelogram Rule, which we can state as follows.

Parallelogram Rule. To find the tangent to a curve at a point F , first decompose the motion of F along the curve into two or more independent motions for which the instantaneous velocity can be found; then, use the parallelogram law of velocities to find the resultant, which lies along the tangent line.

Unfortunately, it proved all too easy to apply the Parallelogram Rule incorrectly. Many historians believe that Roberval himself did so. For example, in *The Historical Development of the Calculus*, C. H. Edwards discussed in detail the case of the ellipse (as well as the parabola and the cycloid) and concluded that “it was something of a

stroke of good fortune that Roberval obtained the correct tangent lines to the parabola and ellipse by this method.” [7, p. 137] More generally, Roberval’s biographer, L. Auger, remarked that

This method is far from being general and . . . the cases where it does not succeed are more numerous than those where it is crowned with success As a matter of fact, the cases where Roberval applied his method were not blemished with error, whether he chose them or was favored by luck [2, p. 63]

Criticism of this argument goes back to the nineteenth-century mathematician J. Duhamel, who stated that Roberval

gave false rules for the determination of tangents to curves generated by radius vectors directed toward fixed centers. He applied these rules in particular cases where they succeeded. [6, pp. 257–258]

Duhamel’s point was that the Parallelogram Rule assumes the independence of the component motions. He specifically criticized its use for curves given in bipolar coordinates, showing that the rule is correct only if either the position vectors are orthogonal or the two velocity vectors are equal in magnitude. In the ellipse, therefore, the Parallelogram Rule does not correctly determine the tangent vector, although the diagonal of the completed parallelogram happens to lie along the correct tangent line.

But was Roberval unaware of the restrictions on the Parallelogram Rule? Roberval’s illustration for the ellipse, unlike Figure 2, did *not* show a parallelogram; the points D' and I' and the lines $D'I'$ and CI' have been added to the original diagram. Roberval’s instructions were simply to bisect $\angle AFC$ or $\angle BFD$ by IFM .

Another “stroke of good fortune”? Before deciding, consider a further example of his work, the quadratrix (Figure 3). Roberval used the ancient definition of this curve as the locus of points F of intersection of two lines, one (FI) descending with uniform speed through parallels to AB , the other (FA) rotating with uniform speed about a center A . One might be tempted to apply the Parallelogram Rule to the vertical vector together with the vector tangent to the circle with radius FA . In a manuscript published posthumously [3], we can see Descartes using just this construction. Roberval, however, saw that the Parallelogram Rule would be misapplied here, because the motion of the point of intersection is not

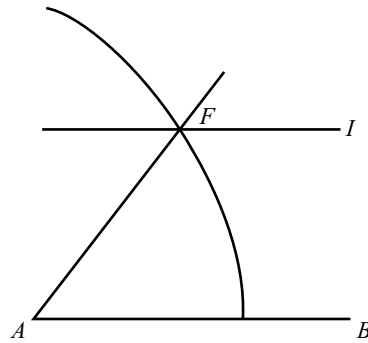


Figure 3.

determined by these two motions. He noted that in addition to the two movements already mentioned, the point F has a further one that obliges it to remain in the intersection of FI with FA . He says

Because these two movements are not the only ones, I do not draw from R a line parallel to FI and equal to FK , to have at its other end the point of the tangent, but rather look at all the movements of F which describe the quadratrix. [14, p. 62]

Andersen [1, p. 298] suggests that Roberval, aware of the limitations of the Parallelogram Rule, had developed a new kinematic method. She has rewritten Roberval’s argument using the language of vectors, a valuable service since it is hard to see Roberval’s method in the original turbid exposition.

Some of our difficulties in understanding Roberval’s exact thoughts about the quadratrix and the ellipse arise because the text of Roberval’s *Observations* consists of the notes of his student de Verdus, who evidently understood very little of what he was writing. In discussing the Archimedean spiral, for example, the text devotes a full page to saying that at constant angular speed, the linear speed of rotation varies from point to point with the instantaneous radius. Yet Roberval was fully able to deal with a variety of curves where the component motions are not uniform. Again, Andersen remarks that, “Verdus described Descartes’s parabola incorrectly, but nevertheless found the right tangent” [11, p. 179, n. 8]. The notes, therefore, do not adequately represent the subtlety of Roberval’s thought. We have Roberval’s word for this; he inserted several deprecating marginal notes in de Verdus’s text. For example, Roberval notes of one explanation: “badly applied, but easy to understand.” Of the determination of the tangent to the quadratrix, Roberval writes, “This proposition is too long and embroiled.” In-

deed, it reads as if de Verdus, unsure where the argument lay, wrote down everything he could imagine to be relevant. It seems safest, therefore, to assign any infelicities and mistakes to de Verdus, as Roberval always obtained the correct tangents.

To exculpate Roberval, however, is not to assert that his lectures were ever expressed with perfect clarity. *We* think of composition of motions in terms of vectors from the start: both the position and the velocity are given by vectors. Roberval, however, was one of those who were just developing that language. He was able to express clearly the vector sum of two velocities, but not the sum of two position vectors, themselves functions of time. Because of this, the composition of motions might mean different things in different cases. Thus Roberval could not easily say why the quadratrix was composed of more than just the circular and the vertical motion. Two decades later, Newton struggled with the same difficulties and emerged triumphant. Thanks to Whiteside's wonderful edition of Newton's mathematical papers, we are able to trace his thoughts.

3 What Newton thought

Newton's investigations of the kinematic method of tangents seem to echo those of Roberval. Did Newton know the latter's work? One possible conduit might have been Isaac Barrow. Barrow himself had investigated the kinematic method about a year and a half before Newton. Then he wrote to Collins:

If you remember, Mersennus and Torricellius do mention a general method of finding the tangents to curve lines by composition of motions, but do not tell it us. Such a one I have something found out. [*sic*] [13, p. 34]

He reported his own method in the first portions of his *Geometrical Lectures* (published 1670), where he used it principally not to *find* tangents but to prove qualitative theorems about them. He did, however, show, "by way of specimen," how his method could be used to determine the tangents of "Cycloids, and curves described like them."

We do not know how much Barrow influenced Newton. We do know that, at the beginning of his mathematical studies, Newton read an edition of Descartes's *Geometry* that included a detailed commentary by Schooten concerning the cycloid. This, in addition to a more general passage in Descartes's *Principles of Philosophy*, may have introduced Newton to the basic idea of composition of motions.

Whatever he may have read or heard, the records show that Newton worked out the kinematic method of tangents for himself.

We can trace his thoughts in a series of manuscripts written between October 1665 and October 1666. These papers (all of which appear in [16]) comprise

1. notes written in October 1665, entitled "How to draw tangents to Mechanichall lines,"
2. a revision of the previous notes, given the same title, dated 8 November 1665,
3. a more complete paper dated 13 November 1665,
4. two drafts of his revised thoughts on limit-motion, 14 and 16 May 1666,
5. the October 1666 tract on fluxions.

In them, Newton discussed the problem of finding tangents to several of the same curves that Roberval had considered, including the cycloid, the ellipse, and the quadratrix. In the first paper, Newton used the Parallelogram Rule. His application of this rule to the quadratrix incorrectly determined its tangent (Figure 4) (and only fortuitously determined the correct tangent line to the ellipse). Within a few weeks, however, Newton had seen the error of applying the Parallelogram Rule in these cases, and he corrected the examples in the succeeding papers. In his November 13 paper, Newton first wrote:

Find . . . in w^t proportion those two lines to w^{ch}
 y^e crooked line is chiefly related doe increase

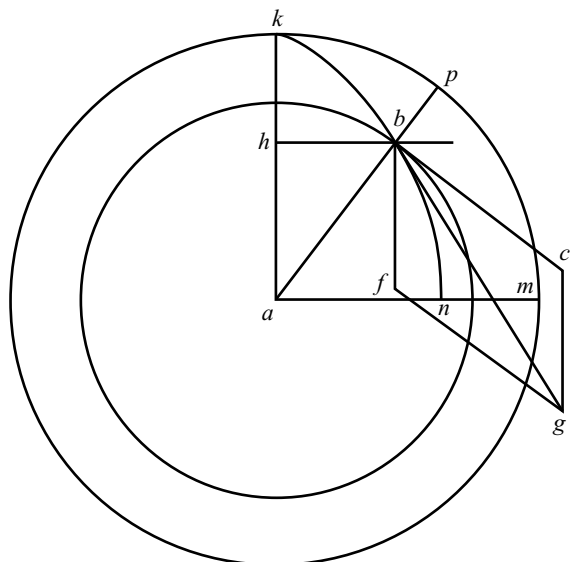


Figure 4.

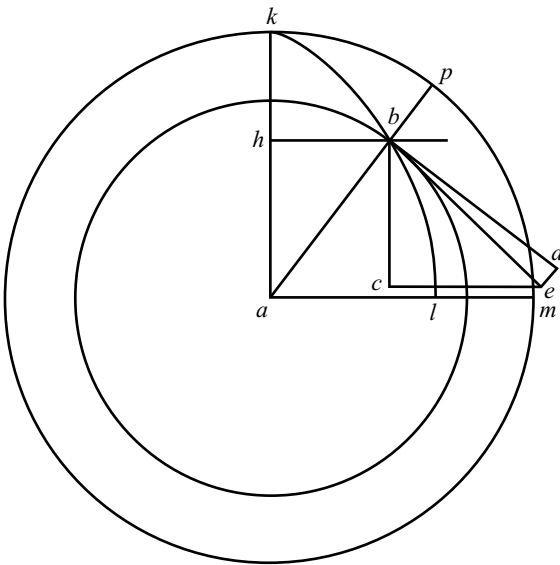


Figure 5.

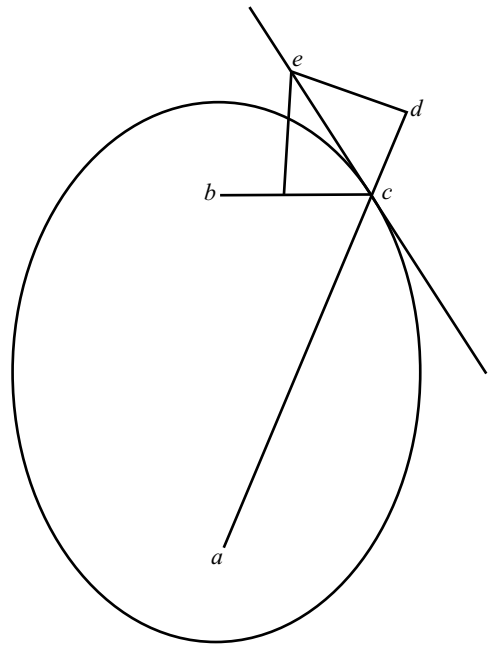


Figure 6.

or decrease; produce y^m in y^t proportion from y^e given point in y^e crooked line; at those ends draw *perpendiculars to y^m* , through whose intersection y^e tangent shall passe. [16, p. 386, my italics]

This seems to describe Newton's treatment of the curve ("crooked line") of the quadratrix (Figure 5) and also of the ellipse (Figure 6). Why should one look for the intersection of the perpendiculars to the tangents to the component motions? Perhaps Newton was still thinking of particular examples where the two motions were instantaneously perpendicular. Or maybe Newton's idea was that an infinitesimal motion displaced the point onto a small circle centered on the original point.

Whatever his idea, he had second thoughts as he wrote the manuscript, so that he replaced the italicized words in the preceding quotation with these: "lines in which those ends are inclined to move." The change did not affect the examples of the quadratrix or the ellipse, since in those cases, the points are inclined to move along the perpendiculars. (Newton explicitly said this of the ellipse.) Nevertheless, his correction is significant, because it shows that by this time, Newton had obtained a basic geometrical insight *via* kinematics: the figure that determines the tangent vector is a quadrilateral, but not necessarily a parallelogram. In the papers that followed, he systematized this, his third method, and incorporated it into more general settings.

Propositions Six and Seven from the May 14, 1666 manuscripts (Figures 7 and 8) show the principles that Newton saw behind his method.

Prop 6^t. If y^e streight line ea doth rest & da doth move: so y^t y^e point a fixed in y^e line da moveth towards it: Then from y^e moving line da drawing $de \parallel ab$, & y^e same way w^{ch} y^e point a moveth; These motions, viz[:] of y^e fixed point a towards b , of y^e intersection point a in y^e line ad towards d , & of y^e intersection point a in y^e line ae towards e , shall bee to one another, as their correspondent lines de , ad , & ae are.

Prop 7^t. If y^e streight lines adm , ane , doe move, soe y^t y^e point a fixed in y^e line amd moveth towards b , & y^e point a fixed in y^e line

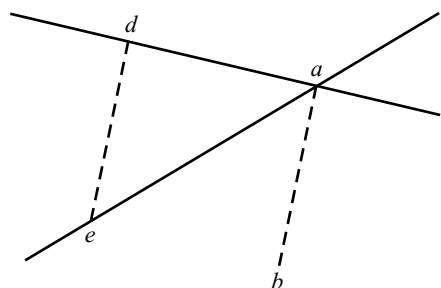


Figure 7.

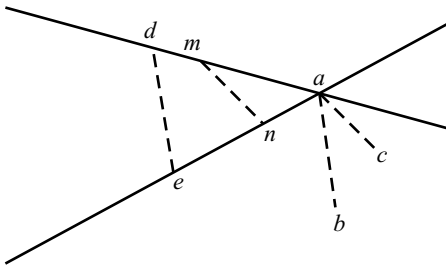


Figure 8.

ae moveth towards c : Then from the line amd , draw $de \parallel ab$ & y^e same way; & from y^e line ae draw $nm \parallel ac$, & y^e contrary way, to make up y^e Trapezium $denm$. And if any two of these four lines de , mn , md , ne , bee to any correspondent two of these four motions, viz: of y^e points a (fixed in y^e line dma) towards b , of y^e point a (fixed in y^e line ane) towards c , of y^e intersection point a moving in y^e line dma according to y^e order of y^e letters m , d , & of y^e intersection point a in y^e line ane according to y^e order of y^e letters n , e : Also all y^e four lines shall be one to another as those four motions are. [16, p. 391]

Proposition Six expresses a principle that we can easily translate into vector language: $\vec{ab} = \vec{ae} + \vec{af}$, where the three vectors represent the displacement of the point a in the plane, the displacement of the point of intersection in the fixed line ae , and the opposite of the displacement of the point of intersection in the moving line ad . Proposition 7 discusses the situation when both lines move. It expresses the movement of the point of intersection in two ways, as movement *in* line md plus movement *of* line md , and as movement *in* line ne plus movement *of* line ne . If one knows the movements of the lines, one can draw a quadrilateral, and its diagonal (from a) represents the motion of the point a in the plane. Of course, Newton was chiefly interested in the cases where the lines are curved; in that case he used the tangent lines to determine the infinitesimal motion. When Newton rewrote the preceding principles two days later, he stated explicitly that the lines might be curved.

Here Newton articulated the basis for the quadrilateral construction that had first appeared in his November 13 manuscript. A rationale for a correct kinematic determination of tangents, inchoate in the lecture notes of Roberval and unclear to Newton half a year earlier, was now revealed. Newton's method

was *not* a simple vector decomposition of velocities, but rather the inference drawn from such decompositions with respect to two different moving frames.

We begin to understand the significance of Newton's discovery when we look back at his paper. Newton had begun with the problem of drawing tangents to *mechanical* curves, but the quotation from the November 13 paper actually refers to the problem of determining tangents to "Geometricall lines"—that is, to algebraic curves. Descartes, who had introduced the distinction in his *Géométrie*, had defined geometric curves in terms of compositions of simple motions. Thus a geometric curve is traced by a point that moves in such a way that two distances, x and y , measured from the point, are related algebraically. Before writing the November 13 paper, Newton had already shown how to calculate the corresponding algebraic relation between the velocities (dx/dt and dy/dt , as we would say). In this way, he could determine, at any point, the ratio of lengths of the velocity vectors of the component motions. With the kinematic method, he could then determine the tangent vector to the curve traced by the two motions. For example, in an ellipse (Figure 6), he let the distances from the foci be x and y , respectively. Then $x + y = c$, a constant, and (as we would say) $dx/dt = -dy/dt$. Newton therefore laid off equal lengths along the two radii, but (since the velocities have opposite sign) one directed away from its focus, the other toward its focus. He completed the quadrilateral by drawing perpendiculars at the ends of these vectors (because those lines represent the directions of motion of the radii). The intersection of these perpendiculars determined the fourth vertex of a quadrilateral based at the point on the curve. The diagonal through the point on the curve and the fourth vertex was the tangent.

4 Conclusion

Since Newton's kinematic method is largely ignored today, one might guess that it exerted little influence on the development of the calculus. On the contrary, Newton made it a major part of his new methods. Any coordinate system in the plane, after all, determines two families of coordinate curves, and the intersection of a curve from each family determines a point. Given a curve with an algebraic equation in these coordinates, the technique of the preceding example could be used to determine its tangents. Newton showed the general utility of the technique in his

October 1666 tract on fluxions, a work that summarized his researches up to that time. The kinematic method had entered the mainstream of the developing calculus.

References

1. K. Andersen, Precalculus, 1635–1665, *Companion Encyclopedia of the History and Philosophy of the Mathematical Sciences*, ed. I. Grattan-Guinness, Routledge, London and New York, 1994, pp. 292–307.
2. L. Auger, *Un savant méconnu: Gilles Personne de Roberval*, Librairie Scientifique A. Blanchard, Paris, 1962.
3. I. Barrow, *The Geometrical Lectures of Isaac Barrow*, ed. J. Child, Open Court, Chicago and London, 1916.
4. R. Descartes, *Les principes de philosophie*, Chez Jean-Baptiste Besongne, Rouen, 1698.
5. —, Excerpta es MSS. R. Des-Cartes: Tangens Quadrataria, *R. Des-Cartes Opuscula Posthuma, Physica et Mathematica*, Amsterdam, 1701.
6. J. Duhamel, Note sur la méthode des tangentes de Roberval, *Memoires présentés par divers savants à l'Académie des sciences de l'Institut de France*, 5 (1838) 257–266.
7. C. H. Edwards, *The Historical Development of the Calculus*, Springer-Verlag, New York, 1979.
8. Galileo Galilei, *Dialogues Concerning Two New Sciences*, tr. H. Crew and A. de Salvio, McGraw-Hill, New York, Toronto, London, 1914.
9. A. Malet, *From Indivisibles to Infinitesimals*, Bel-laterra, Universitat Autònoma de Barcelona, 1996.
10. I. Newton, *The Mathematical Papers of Isaac Newton*, ed. D. T. Whiteside, vol. 1, Cambridge University Press, Cambridge, 1967.
11. K. Pedersen, Roberval's Method of Tangents, *Centaureus*, 13 (1968) 151–182.
12. K. Pedersen, Techniques of the Calculus, 1630–1660, *From the Calculus to Set Theory, 1630–1910*, ed. I. Grattan-Guinness, Duckworth, London, 1980.
13. S. R. Rigaud, *Correspondence of Scientific Men of the Seventeenth Century*, Oxford University Press, Oxford, 1841.
14. G. P. de Roberval, Observations sur la composition des Mouvmens, et sur le moyen de trouver les Touchantes des lignes courbes, *Memoires de l'Académie royale des Sciences depuis 1666 jusqu'à 1699*, VI (1730) 1–89.
15. E. Torricelli, *Opere di Evengelista Torricelli*, ed. G. Loria and G. Vassara, vol. II, Faenza, 1919.
16. D. T. Whiteside, Patterns of Mathematical Thought in the later Seventeenth Century, *Archive for History of the Exact Sciences*, 1 (1961) 179–388.

On the Discovery of the Logarithmic Series and Its Development in England up to Cotes

JOSEF EHRENFRIED HOFMANN

Mathematics Magazine 14 (1939), 37–45

To the expert of today the logarithmic series appears to be a very non-essential detail. In its time it was a very notable discovery as regards itself alone, as well as in the framework of the general theory of series. It was discovered *circa* 1667 by Newton and independently by Mercator. Huygens and Gregory were close to the same discovery but they were anticipated by the other two. Newton was then 24 years old, Mercator 47. For Newton the logarithmic series was a beginning, for Mercator the climax.

1 Nicolas Mercator (1620–1687)

Mercator's life work is almost forgotten today, certainly unjustly. Mercator was a distinguished mathematician, physicist and astronomer. Shortly after his arrival in London the much-traveled man was received into the Royal Society. Products of that period are his new astronomical theory [10], the edition of Euclid [11], the navigation problems [12] and the calculation of logarithms [13]. We shall be concerned here with the latter.

The *Logarithmo-technia* is divided into three very unequal parts. The first two sections, which had already been published separately in 1667, are devoted entirely to the calculation of a system of common logarithms. In the presentation of logarithms Mercator proceeds very intuitively and clearly according to the then generally customary usage. He divides the number domain between 1 and 10 by insertion of geometric means (he calls them *ratiunculae*) into 10 million parts. Thus the logarithm of a number between 1 and 10 is determined from the

number of *ratiunculae* between 1 and this number. Mercator now develops his process for the calculation of the common logarithm of two bases (as such he chooses, stated in modern form, 1.005 and 0.995). It is very carefully thought out and in contrast to the previously adopted methods of calculation has the great advantage that the calculation is purely rational. By continued squaring Mercator gets two successive second powers of the base, between which is the number 10. Now the smaller of these powers is multiplied by the descending series of the previous squares until the new power exceeds 10. Then the performance is repeated until one has the two successive integral powers of the base, between which lies the number 10. Now several further decimal places of the power of the base are determined, which becomes approximately equal to 10, by means of the *regula falsi*. Hence there follows by division the number of the *ratiunculae* which are referred to the base, i.e., its logarithm.¹

Now Mercator introduces the absolute value of the logarithm from the ratio of two positive magnitudes as its "proportional measure" and teaches calculation with these proportional measures. Then he forms logarithms from the successive members of an arithmetical series and shows that the terms of their difference series become smaller and smaller. He builds up the logarithms themselves from the first terms of

¹If we set, e.g. $1.005 = g$, then Mercator forms $g^2, g^4, g^8, g^{16}, g^{32}, g^{64}, g^{128}, g^{256}$, and finds that $g^{512} > 10$. Hence he reckons further $g^{256+128}, g^{384+64}, g^{448+32} > 10$, $g^{448+16} > 10$, $g^{448+8}, g^{456+4}, g^{460+2} > 10$, $g^{460+1} < 10$. From $g^{461} = 9.965774$ and $g^{462} = 10.015603$ he finds $g^{461.6868} \approx 10$; therefore the number of *ratiunculae* pertaining to g is ten million: $461.6868 = 21,659.7$. Consequently the common logarithm of $g = 0.00216597$ (instead of 0.00216606).

the difference series. The logarithms of powers and roots are approximated by means of the logarithms of rational approximation values. There is attached an excellent process for the gradual refinement of the calculation. As a supplement several formulas are given which serve convenience in calculation. Then follow more exact directions for the practical calculation of a complete table of logarithms.

In the third section the ordinate

$$y = \frac{1}{1+x} = 1 - x + x^2 - x^3 \dots$$

of the equilateral hyperbola is transformed, by dividing out, into a power series. The surface of the hyperbola segment is built up entirely in the sense of Cavalieri's method of infinitesimals from the totality of all parallel coordinates "contained" in it. How one is to take and combine the sums over the single powers of x is only briefly alluded to. [From the detailed explanations, it seems to follow that Mercator had not studied Cavalieri's original works but had learned from lectures heard in Rostock, Copenhagen and Danzig.] Thus Mercator gets the hyperbolic segment in that form which we would today write thus:

$$\int_0^x \frac{dt}{1+t} = x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} \dots$$

However he does not express his result by a formula, but entirely in words. By means of an extremely bold conclusion he finds that this series can also be expressed by the logarithm of $(1+x)$. The paper ends with the calculation of the body which "consists" of infinitely many hyperbolic segments. We would thus write the result expressed again only in words:

$$\int_0^x \log(1+t) dt = \frac{x^2}{1 \cdot 2} - \frac{x^3}{2 \cdot 3} + \frac{x^4}{3 \cdot 4} - \frac{x^5}{4 \cdot 5} \dots$$

This third section of the *Logarithmotechnia* was at once announced by John Wallis through a review in the *Phil. Trans.* [17], in which the formal notation is very much improved. Wallis calls attention to the fact that Mercator's development is admissible only for $x < 1$ and adds the development of

$$\int_0^x \frac{dt}{1-t} = x + \frac{x^2}{2} + \frac{x^3}{3} + \frac{x^4}{4} + \dots \quad (0 < x < 1)$$

In a note by Mercator himself [9, I, 227–232] the logarithms determined from the hyperbolic segments are expressly designated as "natural logarithms". Here the values of $\log 2$, $\log 3$, $\log 10$, and $\log 11$

are determined from the correctly combined series for

$$\log(1+x) \text{ and } \log \frac{1}{1-x}, \text{ with } x = 0.1 \text{ and } 0.2.$$

Next Mercator by means of multiplication with $0.43429 = 1/\log 10$ transforms from the natural to the common logarithms and *vice versa*.

2 James Gregory (1638–1675)

A few months later Gregory came before the public with his *Exercitationes geometricae* (London, 1668), whose interesting preliminary history I must briefly go into. On account of his *Vera circuli et hyperbolae quadratura* (Padua, 1667), in which he sought to prove the algebraic impossibility of squaring a circle, Gregory had involved himself in a heated quarrel with Huygens. The very adverse, indeed unjust criticism of Huygens in the *Journal des Sçavans* (July, 1668) was followed by a rather irritable reply by Gregory in the *Phil. Trans.* (July, 1668), then a reply, unbearable for Gregory, by Huygens in the *Journal des Sçavans* (Nov. 1668) and a brisk correspondence between Huygens and the most influential members of the Royal Society, who would gladly have smoothed out the affair and thus hindered a further explanation by Gregory. The latter fell into frightful excitement. He believed he was persecuted, disliked and slighted on all sides. The *Exercitationes geometricae*, on which he was just then working and in which he took a stand on numerous questions of the day, were misused in the foreword for a huge counter-attack against Huygens [9, II, 2–5]. Later the quarrel was put aside by the Royal Society—one might almost say, by compulsion—and Gregory's book was rather hushed up on account of the malicious introduction. This fate was undeserved; for the mathematical content of the *Exercitationes* is important.

For us only a small part of the little work is important [7]. There Gregory, who depends on the geometric integration method of Gregoire de Saint Vincent [4] and most probably was familiar with Fermat's essay *De aequationem localium transmutatione, & emendatione* . . . [3], gives a completely impeccable proof for the representation of the hyperbolic segment by means of the power series. Moreover he combines two adjacent segments and forms that series which we represent today by

$$\frac{1}{2} \log \frac{1+x}{1-x}$$

but everything is still expressed very ponderously. The relation to logarithms is only cursorily touched on and dismissed rather superficially. Meanwhile, we have a letter of Gregory of a later date, which gives us a better insight into his accomplishments [15, 240]. [The letter is dated April 9, 1672 and addressed to the weights and measures officer Michael Dary, who rated as a good algebraist and was counted among the more intimate friends of John Collins. Unfortunately the original is lost; we must rely on a copy by Collins in which the nomenclature was probably altered from that of the original.] Indeed there a subdominant (Minorante) is set up for the logarithmic series. It arises by setting from any member of the series for

$$\frac{1}{2} \log \frac{1+x}{1-x}$$

on, instead of the correct terms of the series those of a geometric series.

3 Isaac Newton (1642–1727)

The logarithmic series plays a great role in the earliest researches of Newton, concerning which we are only very meagerly informed. Fortunately the evolution of this detail can be surveyed rather accurately. Newton probably was occupied in the years 1665 and 1666 with the quadrature of the hyperbolic segment.

By a purely numerical attack which corresponds to the series of Mercator and Wallis he arrived at the surface segments. However, from the beginning on he calculated this from half the sum and difference, and was thus in possession of Gregory's results. He also knew the relation to logarithms. [Probably he learned it through his teacher Isaac Barrow, who even at that time knew that the hyperbolic segment is proportional to the logarithm of the quotient of the including ordinates.] In the *Analysis per aequationes numero terminorum infinitas* [18, 3–28] and in the *Methodus fluxionum et serierum infinitarum cum ejusdem applicatione ad curvarum geometriam* [18, 29–140] the presentation gradually becomes more general and in a letter of January 19, 1689 to Collins [15, 285–286] there appears not only the series for the hyperbolic segment represented by

$$\log \frac{a+x}{a-x}$$

but also the remark that this series converges twice as rapidly as the original series of Mercator.

In the *Analysis per aequationes* the logarithmic series moreover is cleverly gotten by reversion by means of gradual approximation. In the first letter to Oldenburg for Leibniz of June 13, 1676 [16, (2) 32–41], this reversion function is written in quite general form and in the second letter to Oldenburg for Leibniz of October 24, 1676 [16, (2) 130–149] it is gotten out by a sort of comparison of powers. Moreover Newton gives in the *Methodus fluxionum* a process, not yet entirely mature, of finding the logarithm of a from the already known logarithms of $a \pm x$ and the given x .

4 The methodological expansion

With these developments the formulation of the fundamental material, namely the setting up of the logarithmic series and its reversion, is completely closed. That which follows is refinements in details and methodological improvements. They were not immediately successful, indeed not until a full quarter of a century after the first discoveries; at a time therefore, when the new thoughts were no longer so strange and unusual.

Foremost is Edmund Halley (1656–1742) with his attempt to eliminate the hyperbolic surface — that painful transition structure between the series and the logarithm [5], [8].

He explains the logarithmic series and its reversion from the binomial theorem for infinitely small *resp.* infinitely large exponents. In modern terms his process reduces to the change of limits:

$$\log(1+x) = y = \lim_{n \rightarrow \infty} \frac{(1+x)^{1/n} - 1}{1/n}$$

and

$$(1+x) = \lim_{n \rightarrow \infty} \left(1 + \frac{y}{n}\right)^n.$$

It is expressed however without the change of limits, only with $n = \infty$ and in very obscure form. Doubtless Halley first set $n = \infty$ in the binomial series, then established the connection with the logarithmic series and accordingly attempted to get it by reversion. But this was only very slightly successful; it remained in a changed dress with fine but obscure words.

A few years later Abraham de Moivre (1667–1754) explains in a short essay the attack on the undetermined coefficients in the transformation of series — at that time no longer much of which was

new as to thought content [14]. In it the deduction of the logarithmic series is also touched on. It occurs thus:

If $1+z$ is a number and its logarithm is expressed by the series $az+bz^2+cz^3+\dots$, then the logarithm $ay+by^2+cy^3+\dots$ belongs to $1+y$. If now $1+z=(1+y)^n$, therefore

$$az + bz^2 + cz^3 + \dots = nay + nby^2 + ncy^3 + \dots,$$

then one can develop and insert $z=(1+y)^n-1$ according to the binomial theorem. Then the a, b, c , etc. can be determined by comparison of coefficients with the exception of the first a , which remains arbitrary and characterizes the different kinds of logarithms. That is really remarkable enough; for the n which remained arbitrary falls completely out of the calculation. Unfortunately this is not expressly emphasized by de Moivre, although he doubtless knew it. Fundamentally de Moivre attacked the problem as it is done with a functional equation. The passage under consideration is certainly one of the earliest examples of it.

Twelve years later Roger Cotes (1682–1716) again takes up the same thought. He recognizes its deeper meaning and thinks it through clear to the end [2].

The result is the new definition of the logarithm from the functional equation $f(a^x) = xf(a)$. Cotes doesn't yet have our functional signs. He replaces it by a designation which he knew basically since Mercator: $f(a)$ means for him the measure of the relationship a^x . The functional equation is now solved by an infinitesimal method which is equivalent to our treatment by means of differential calculus. Hence there follows

$$f\left(\frac{c+x}{c-x}\right) = M\left(\frac{x}{c} + \frac{x^3}{3c^3} + \frac{x^5}{5c^5} \dots\right)$$

and the modulus M characterizes the various logarithmic systems which can be gotten in such a manner. By reversion of the logarithmic series Cotes on this occasion comes to a continued-fraction development of e (doubtless based on direct dividing out of his result calculated to twelve decimal places) and finally gives the quadrature of the hyperbolic segment and the hyperbolic sector by means of logarithms. These calculations of his we can designate in modern style and very aptly as "logarithmic integration"; that is the essence of the *Logometria*. It is a new method — we would designate it as an example of the substitution method in the transformation of indefinite integrals — by means of which an

abundance of contemporary but apparently mutually unrelated separate results could be explained from a single guiding viewpoint. Cotes himself provided for the best in this respect in the *Logometria* and the fragments of the *Harmonia mensurorum* which he completed; Smith added nothing new of his own.

If we look back then we recognize in this entire development from Mercator to Cotes a coherent line. The thought content becomes piecemeal richer and richer, the form better and better, more and more complete. Cotes stands as the last on the shoulders of all his predecessors. He really gives something finished and complete.

We must add that Leibniz in 1673 worked through Mercator's *Logarithmotechnia* and at least after 1676 was in possession of the reversion of the logarithmic series. However the exact details cannot be presented at the moment as long as the Leibniz edition is not completed; for it will certainly bring new material which remained hitherto inaccessible. Meanwhile, the activity of Leibniz and his school had no further influence on the development of the logarithmic series in England; hence it could be ignored without essential loss.

To the reader of today much in the conception and mode of expression of that time appears strange and unusual. Between us and the mathematicians of the late seventeenth century stands Leonhard Euler (1707-1783). He is the real founder of our modern conception. However non-rigorous he may be in details, he ends and conquers the previous epoch of direct geometric infinitesimal considerations and introduces the period of mathematical analysis according to form and content. Whatever was written after him on the logarithmic series is necessarily based no longer on the already obscured predecessors in the receding mathematical Renaissance, but on Euler's *Introductio in analysin infinitorum*, Lausanne 1748, in which the entire seventh chapter treats of logarithms.

References

1. Henry Briggs, *Arithmetica logarithmica sive logarithmorum chiliadis triginta*, London, 1624.
2. Roger Cotes, *Logometria*, *Philosophical Transactions* 29 (1714), 5–45. Reprinted in Roger Cotes, *Harmonia mensurarum, sive Analysis et Synthesis per rationum et angulorum mensuras promotae; accedunt alia Opuscula mathematica*, Robert Smith, ed. Cambridge, 1722, 1–41.
3. Pierre de Fermat, *De aequationum localium transmutatione, et emendatione ad multimodam curvilinearum*

- inter se vel cum rectilineis comparationem, cui annectitur proportionis geometricae in quadrandis infinitis parabolis et hyperbolis usus, in Clément-Samuel de Fermat, *Varia opera mathematica*, Toulouse: Joannis Pech, 1679, 44–57.
4. Gregory of St. Vincent, *Opus geometricum quadraturae circuli et sectionum conii*, Antwerp, 1647.
 5. Edmond Halley, A most compendious and facile Method for constructing the Logarithms, exemplified and demonstrated from the Nature of Numbers, without any regard to the Hyperbola, with a speedy Method for finding the Number from the Logarithm given, *Philosophical Transactions* 19 (1695), 58–67. Reprinted in [9, II, 84–91].
 6. J. E. Hofmann, Nicolaus Mercators Logarithmotechnia (1668), *Deutsche Mathematik* 3 (1938), 445–466.
 7. ———, Weiterbildung der logarithmischen Reihe Mercators in England, I, *Deutsche Mathematik* 3 (1938), 598–605.
 8. ———, Weiterbildung der logarithmischen Reihe Mercators in England, II, *Deutsche Mathematik* 4, 1939.
 9. Francis Masères, *Scriptores logarithmici*, London, 1791.
 10. Nicolas Mercator, *Hypothesis Astronomica nova, ejusque cum Observationibus consensus*, London, 1664, extended in the *Institutiones astronomicae*, London, 1676.
 11. ———, *Euclidis Elementa Geometrica, novo ordine ac methodo fare demonstrata*, London, 1665. The second edition is augmented by an *Introductio brevis in Geometriam*, London, 1678.
 12. ———, Problemata quaedam, ad promotionem scientiae navigatoriae facientia, *Philosophical Transactions* 2 (1666), 161–163.
 13. ———, *Logarithmo-technia; sive methodus construendi logarithmos nova, accurata, et facilis; scripto antehac communicata, anno sc. 1667, nonis Augusti. Cui nunc accedit vera quadratura hyperbolae, et inventio summae logarithmorum*, London, 1668. A reprint is in [9, I, 169–196]. See also [6].
 14. Abraham de Moivre, A Method of extracting the Root of an Infinite Equation, *Philosophical Transactions* 20 (1698), 190–192.
 15. Stephen Jordan Rigaud, *Correspondence of Scientific Men of the XVIIth Century*, Oxford, 1841.
 16. H. W. Turnbull, et. al. (eds.) *The Correspondence of Isaac Newton*. Cambridge: Cambridge University Press, 1959–78. 7 volumes.
 17. John Wallis, Review of [13], *Philosophical Transactions* 3 (1668), 753–759; reprinted in [9, I, 219–226].
 18. D. T. Whiteside, ed., *The Mathematical Works of Isaac Newton*. New York: Johnson Reprint Co., 1964. 2 volumes.

Isaac Newton: Man, Myth, and Mathematics

V. FREDERICK RICKEY

College Mathematics Journal 18 (1987), 362–389

Three hundred years ago, in 1687, the most famous scientific work of all time, the *Philosophiæ Naturalis Principia Mathematica* of Isaac Newton, was published. Fifty years earlier, in 1637, a work which had considerable influence on Newton, the *Discours de la Méthode*, with its famous appendix, *La Géométrie*, was published by René Descartes. It is fitting that we celebrate these anniversaries by sketching the lives and outlining the works of Newton and Descartes.

In the past several decades, historians of science have arranged the chaotic bulk of Newton manuscripts into a coherent whole and presented it to us in numerous high quality books and papers. Foremost among these historians is Derek T. Whiteside, of Cambridge, whose eight magnificent volumes overflowing with erudite commentary have brought Newton to life again.

By unanimous agreement, the *Mathematical Papers [of Isaac Newton]* is the premier edition of scientific papers. It establishes a new criterion of excellence. Every further edition of scientific papers must now measure itself by its standard. [26 p. 87]

Other purposes of this article are to dispel some myths about Newton—for much of what we previously “knew” about him is myth—and to encourage the reader to look inside these volumes and to read Newton’s own words, for that is the only way to appreciate the majesty of his intellect.

1 Newton’s education and public life

Isaac Newton was born prematurely on Christmas Day 1642 (O.S.), the “same” year Galileo (1564–

1642) died, in the family manor house at Woolsthorpe, some 90 km NNW of Cambridge. His illiterate father—a “wild, extravagant, and weak man”—had died the previous October. His barely literate mother, Hanna, married the Reverend Barnabas Smith three years later, leaving Newton to be raised by his aged grandmother Ayscough.

Newton attended local schools and then, at age 12, traveled 11 km north to the town of Grantham, where he lived with the local apothecary and his books while attending grammar school. The town library had two or three hundred books, some 85 of which are still chained to the walls. Of course he studied Latin, also some Greek and Hebrew. Four years later, in 1658, he returned home to help his now twice-widowed mother manage the farm. Recognizing that Newton was an absent-minded farmer, his uncle William Ayscough (M.A. Cambridge, 1637) and former Grantham schoolmaster, Henry Stokes, persuaded his mother to send him back to Grantham to prepare for Cambridge. Judging by a mathematical copybook in use at Grantham in the 1650s, Stokes was a most unusual schoolmaster. The copybook contained arithmetic through the extraction of cube roots, surveying, elementary mensuration, plane trigonometry, and elaborate geometric constructions, including the Archimedean bounds for π . This went far beyond anything taught in the universities of the period; consequently, contrary to tradition, Newton had a superior knowledge of mathematics before he went to Cambridge [33, pp. 110–111; 34, p. 101; updating 20, I, p. 3].

In 1661, eighteen-year-old Newton matriculated at Trinity College, the foremost college at Cambridge, as a subsizar (someone who earned his way by performing simple domestic services). This position reflected his wealthy mother’s reluctance to

send him to the university. At that time, Cambridge was little more than a degree mill. Lectures were seldom given. Fellows tutored primarily to augment their income. Although Newton did not finish any of the books from the established curriculum, which consisted mostly of Aristotelian philosophy, he did learn the patterns of rigorous thought from Aristotle's sophisticated philosophical system. A chance encounter with astrology in 1663 led him to the more enlightened "brisk part of the University" that was interested in the work of Descartes [28, p. 90]. The laxity of the university allowed him to spend the last year and a half of his undergraduate studies in the pursuit of mathematics. In 1665, Newton received his B.A. "largely because the university no longer believed in its own curriculum with enough conviction to enforce it." [28, p. 141].

In the summer of 1665, virtually everyone left the university because of the bubonic plague. The next March the university invited its students and Fellows to return for there had been no deaths in six weeks, but by June it was clear that the plague had not left, so the students who had returned left again. The university was able to resume again in the spring of 1667. Newton had left by August 1665 for Woolsthorpe. He returned on 20 March 1666, probably left again in June, but not until he had written his famous May 1666 tract on the calculus. He did not return to Cambridge until late April 1667, having revised the May tract into the October 1666 tract while back on the farm. "For whatever it is worth, the papers do not indicate that anything special happened at Woolsthorpe." [27, p. 116]. Much has been written about these plague years as the *anni mirabiles* of Newton, but the record clearly shows that he wrote the bulk of his mathematical manuscripts on the calculus while he was at Cambridge.

Myth: *At the Woolsthorpe farm, during the plague years, Newton invented the calculus so that he could apply it to celestial mechanics.*

The primary source for the myth [27, p. 110] of Newton's miracle years is this 1718 (unsent?) letter from Newton to Pierre DesMaizeaux:

In the beginning of the year 1665 I found the Method of approximating series & the Rule for reducing any dignity [= power] of any Binomial into such a series. The same year in May I found the method of Tangents . . . , & in November had the direct method of fluxions & the next Year in January had the Theory of Colours &

in May following I had entrance into y^e inverse method of fluxions. And the same year I began to think of gravity extending to y^e orb of the Moon & (having found out how to estimate the force with w^{ch} globe revolving within a sphere presses the surface of the sphere) from Keplers rule . . . I deduced that the forces w^{ch} keep the Planets in their Orbs must [be] reciprocally as the squares of their distances from the centers about wch they revolve: & thereby compared the force requisite to keep the Moon in her Orb with the force of gravity at the surface of the earth, & found them answer pretty nearly. All this was in the two plague years of 1665 & 1666. For in those days I was in the prime of my age for invention & minded Mathematicks & Philosophy [= Science] more then at any time since. [27, p. 109]

Lucasian Professor. At Trinity College, Newton became a Minor Fellow in 1667 and a Major Fellow in 1668. On 29 October 1669, at the age of 26, Newton became the second Lucasian Professor of Mathematics at Cambridge, succeeding Isaac Barrow (1630–1677). This post gave him security, intellectual independence, and a good salary. According to the Lucasian statutes, Newton was to lecture once a week during each of the three terms and to deposit ten of the lectures in the library. Even though this position had been designed by its founder Henry Lucas as a teaching post, not a research position [20, V, xiv], Barrow had already turned the position into a sinecure and Newton did not work much harder at the teaching aspects of the post. He deposited 3–10 lectures per year for the first seventeen years as Lucasian Professor, and none thereafter.

As a teacher, Newton left no mark whatsoever. Years later, when he was duly famous, one would expect that many people would have claimed to have attended his lectures, yet we know of only three. Perhaps the situation is best summed up by Newton's amanuensis (a human wordprocessor), the unrelated Humphrey Newton:

He seldom left his chamber except at term time, when he read in the schools as being Lucasianus Professor, where so few went to hear him, and fewer that understood him, that oft-times he did in a manner, for want of hearers, read to the walls. [6, X, 44]

London and Beyond. In 1696, Newton accepted the post of Warden of the Mint (moving to London in



Figure 1. Newton at age 82.

March or April of 1696) and four years later became Master. In 1701, Newton resigned the Lucasian professorship. In 1703, he was elected President of the Royal Society, which he ruled with an iron hand until his death. In 1705, Newton was knighted by Queen Anne — not for his scientific advances, but for the service he had rendered the Crown by running (unsuccessfully) for Parliament in 1705 [28, p. 625]. For the rest of his life, Newton looked after the Mint and the Royal Society, twice revised his *Principia* (1713 and 1726), engaged in the infamous priority dispute with Leibniz, and toiled on secret research in religion and church history. His creative scientific life essentially ended when he left Cambridge.

Newton died 20 March 1727, at the age of 84, having been ill with gout and inflamed lungs for some time. He was buried in Westminster Abbey.

Newton's Nachlass. At the time of his death Newton was wealthy. Income from the Lucasian Chair and farm rents brought £250 per year, sufficient for a handsome living for a bachelor Don. When he became Master of the Mint, his salary jumped to £600 and he also received the perquisite of a commission on the amount of coinage. This amounted to some £1500 per year, thus bringing his income to over £2000 per year, a very substantial figure at that time. On his death his estate was valued at £30,000.

Newton left his library of some two thousand volumes to his nieces and nephews. The books were quickly sold to the Warden of Fleet Prison for £300 for his son Charles Huggins who was a cleric near Oxford. On Huggins' death in 1750 they were sold to his successor, James Musgrave, for £400. They remained in the Musgrave family until 1920, when some of them were sold at auction as part of a "Library of miscellaneous literature", fetching only

£170. Although the family didn't know what they had sold, the book dealers knew what they had bought. Newton's annotated copy of Barrow's *Euclid*, which sold for five shillings, was soon in a bookseller's catalogue for £500. In 1927, the remaining 858 volumes were offered for £30,000 but remained unsold until 1943 when they were purchased for £5,500 and donated to the Wren Library at Trinity College. Of the thousand or so that were dispersed in 1920, some still show up unrecognized in bookshops. As recently as 1975, one was purchased in a Cambridge bookshop for £4. The books are easily identified by Newton's peculiar method of dog-earring by folding a page down to point to the precise word that interested him.

From Newton's library, 1736 books have now been located. Since his was a working library, a subject classification of the non-duplicates provides some information about Newton's interests. (For additional details, see [10, p. 59], from which this table is condensed.)

Subject	No. of Titles	Percentage
Mathematics	126	7.2
Physics/Astronomy	85	4.9
Alchemy	169	9.6
Theology	477	27.2
History	143	8.2
Other Science	158	9.0
Other	594	33.9

Newton also had access to the library of Barrow until Barrow's death in 1677, and to the Cambridge libraries until he moved to London in 1696.

Whiteside has tracked down every available scrap of material on Newton's mathematics and published it in *The Mathematical Papers of Isaac Newton* [20]. To really appreciate Newton's mathematical genius, one must grapple with his mathematics as he wrote it. The best place to gain an overview for this project is in Whiteside's wonderful introductions to these volumes and to the various papers in them. They have been used extensively in preparing this paper.

This biographical sketch has been intentionally kept short. For further details about Newton and his work, see the article by I. B. Cohen in the *Dictionary of Scientific Biography* (DSB) [6, X, pp. 42–103]. This is the single most authoritative reference work about the lives and contributions of deceased scientists. To avoid frequent references to it, we give dates after the first occurrence of an individual's name if the DSB contains an article about him. Two excellent biographies of Newton are Westfall's full sci-

entific biography, *Never at Rest* [28], and Manuel's psychobiography *A Portrait of Isaac Newton* [15], some conclusions of which must be taken with care. For mathematical details, consult the many papers of Whiteside, only a few of which are cited here.

2 Newton's mathematical readings

The year 1664 was a crucial period in Newton's development as a mathematician and scientist, for it was then that he began to extend his readings beyond the traditional Aristotelian texts of the moribund curriculum to the new Cartesian ideas. (For details of Newton's non-mathematical readings, see McGuire [16].) According to Abraham De Moivre (1667–1754), the expatriate French intimate of Newton during Newton's last years, the immediate impulse for Newton taking up mathematics was:

In 63 [Newton] being at Sturbridge [international trade] fair bought a book of Astrology, out of a curiosity to see what there was in it. Read in it till he came to a figure of the heavens which he could not understand for want of being acquainted with Trigonometry.

Bought a book of Trigonometry, but was not able to understand the Demonstrations.

Got Euclid to fit himself for understanding the ground of Trigonometry.

Read only the titles of the propositions, which he found so easy to understand that he wondered how any body would amuse themselves to write any demonstrations of them. Began to change his mind when he read that Parallelograms upon the same base & between the same Parallels are equal, & that other proposition that in a right angled Triangle the square of the Hypothenuse is equal to the squares of the two other sides.

Began again to read Euclid with more attention than he had done before & went through it.

Read Oughtreds [Clavis] which he understood tho not entirely, he having some difficulties about what the Author called Scala secundi & tertii gradus, relating to the solution of quadratick [&] Cubick Equations. Took Descartes's Geometry in hand. tho he had been told it would be very difficult, read some ten pages in it, then stopt, began again, went a little farther than the first time, stopt again, went

back again to the beginning, read on till by degrees he made himself master of the whole, to that degree that he understood Descartes's Geometry better than he had done Euclid.

Read Euclid again & then Descartes's Geometry for a second time. Read next Dr Wallis's *Arithmetica Infinitorum*, & on the occasion of a certain interpolation for the quadrature of the circle, found that admirable Theorem for raising a Binomial to a power given. But before that time, a little after reading Descartes Geometry, wrote many things concerning the vertices Axes [&] diameters of curves, which afterwards gave rise to that excellent tract *de Curvis secundi generis*.

In 65 & 66 began to find the method of Fluxions, and writt several curious problems relating to that method bearing that date which were seen by me above 25 years ago. [20, I, pp. 5–6]

These words of De Moivre, which agree with the report of Conduitt [20, I, pp. 15–19], certainly have an air of authenticity to them, and we know, based on extant manuscripts, that they are substantially correct (modulo Stokes's copybook). In the years 1664–1665, Newton made detailed notes on the following contemporary high level books, which influenced him at the very beginning of his mathematical studies.

- Barrow's *Euclidis Elementorum* (1655)
- Oughtred's *Clavis Mathematicae* (1631) in the 1652 edition
- *Geometria, à Renato des Cartes*, 1659–1661 edition of Schooten
- Schooten's *Exercitationum Mathematicarum Libri Quinque* (1657)
- Viète's *Opera Mathematica*, 1646 edition of Schooten
- Wallis's *Arithmetica Infinitorum* (1655)
- Wallis's *Tractatus Duo* (1659).

Let us look carefully at each of them to see what Newton learned.

Euclid (fl. ca. 295 B.C.). As De Moivre indicated, Newton read Euclid as a student, although he did not develop any deep knowledge of the work then. Recall the story [28, p. 102] that Barrow examined Newton on Euclid and found him wanting. Newton was mainly influenced by books II (geometrical

algebra), V (proportion), VII (number theory), and X (irrationals). The primary thing that he learned from Euclid was the traditional forms of mathematical proof [20, I, p. 12].

William Oughtred (1575–1660). At age fifteen, Oughtred went to Cambridge where he studied mathematics diligently on his own, for there was then hardly anyone there to teach him. He graduated B.A. in 1596 and M.A. in 1600. In 1603, he became a (pitiful) preacher and soon settled in as rector at Albury where he remained until his death.

It was as a teacher that he was renowned. He taught privately and for free. People came from the continent to talk to him, so wide was his reputation in mathematics. To instruct a young Earl, Oughtred wrote a little book of 88 pages that contained the essentials of arithmetic and algebra. *Clavis Mathematicae (Key to Mathematics)* published in 1631, was “a guide for mountain-climbers, and woe unto him who lacked nerve.” [2, p. 29]. The style was obscure, the rules so involved they were difficult to comprehend. Oughtred carried symbolism to excess, a habit acquired by his most famous pupil, John Wallis. Nonetheless, *Clavis* established Oughtred as a capable mathematician and exerted a considerable effect in England, for it was a widely studied book in higher mathematics [32, p. 73].

Oughtred’s *Clavis*, in the 1652 edition, was one of the first mathematical books that Newton read. From it he learned a very important lesson: Oughtred taught that *algebra was a tool for discovery* that did not need to be backed up by geometry [13, p. 408]. Newton held Oughtred in high regard, describing him as “a Man whose judgment (if any man’s) may be safely relied upon.” [19, III, p. 364]

René Descartes (1596–1650). René du Perron Descartes was born 31 March 1596 in La Haye (now La Haye-Descartes), France, a small town 250 km SSW of Paris. At the age of eight, he enrolled in the new Jesuit *collège* at La Flèche. There Descartes received a modern education in mathematics and physics — including the recent telescopic discoveries of Galileo — as well as more traditional schooling in the humanities, philosophy, and the classics. It was there, because of his then delicate health, that he developed the habit of lying abed in the morning in contemplation. Descartes retained an admiration for his teachers at La Flèche but later claimed that he found little of substance in the course of instruction and that only mathematics had given him any certain knowledge.

Descartes graduated in law from the University of Poitiers in 1616, at age 20, but never practiced law as his father wished. By this time, his health improved and he enjoyed moderately good health for the rest of his life. Because he decided that he could not believe in what he had learned at school, he began a ten-year period of wandering about Europe, spending part of the time as a gentleman soldier. It was during this period that Descartes had his first ideas about the “marvelous science” that was to become analytic geometry.

Although we have little detail about this period of his life, we do know that he hoped to learn from “the book of the world.” Descartes reached two conclusions. First, if he was to discover true knowledge he must carry out the whole program himself, just as a perfect work of art is the work of one master. Second, he must begin by methodically doubting everything taught in philosophy and looking for self-evident, certain principles from which to reconstruct all science.

In November 1628, Descartes had a public encounter with Chandoux, who felt that science was founded only on probability. By using his method to distinguish between true scientific knowledge and mere probability, Descartes easily demolished Chandoux. Among those present was the influential Cardinal de Bérulle, who charged Descartes to devote his life to working out the application of “his manner of philosophizing . . . to medicine and mechanics.” To execute this design, Descartes moved to the Netherlands in 1628, where he lived for the next twenty years.

In Holland, Descartes worked at his system and, by 1634, had completed a scientific work entitled *Le Monde*. He immediately suppressed the book when he heard about the recent condemnation of Galileo by the inquisition. He learned this from Marin Mersenne (1588–1648), a fellow student at La Flèche and later the hub of the scientific correspondence network in Europe. This reveals Descartes’ spirit of caution and conciliation toward authority (he was a lifelong devout Catholic). Later he took care to present his less orthodox views more obliquely.

Three hundred and fifty years ago, in 1637, the *Discours de la Méthode* [Figure 2], with appendices *La Dioptrique*, *Les Meteores*, and *La Géométrie*, appeared anonymously in Leyden, although it was soon widely known that Descartes was the author. The opening *Discours* is notable for its autobiographical tone, compressed presentation, and elegant

DISCOURS
DE LA METHODE

Pour bien conduire la raison, & chercher
la verité dans les sciences.

PLUS
LA DIOPTRIQUE.
LES METEORES.
ET
LA GEOMETRIE.

Qui sont des essais de cete METHODE.



A LEYDE
De l'Imprimerie de IAN MAIR F.
M D C XXXVII.
Avec Privilège.

Figure 2. *Discours de la Methode*

French style. It was written in French since he intended — as did Galileo — to aim over the heads of the academic community to reach the educated people. Today, it is this opening *Discours*, with its problem-solving techniques that is read. (Pólya was very much influenced by Descartes. [22, I, p. 56])

Descartes' Rules: The first was never to accept anything as true that I did not know evidently to be such; that is to say, carefully to avoid haste and bias, and to include nothing more in my judgements than that which presented itself to my mind so clearly and so distinctly that I had no occasion to place it in doubt.

The second was to divide each of the difficulties that I examined into as many parts as possible, and according as such division would be required for the better solution of the problems.

The third was to direct my thinking in an orderly way, by beginning with the objects that were simplest and easiest to understand, in order to climb little by little, gradually, to the knowledge of the most complex; and even for this purpose assuming an order among those objects which do not naturally precede each other.

And the last was at all times to make enumerations so complete, and reviews so general, that I would be sure of omitting nothing. [4, p. 16]

In 1644, Descartes published *Principia Philosophiae*, a work in which he presented his views



Figure 3. Rene Descartes

on cosmology. He expounded a mechanical philosophy in which a body could influence only those other bodies that it touched. Thus, for example, Descartes imagined space filled with “vortices” that moved the planets. This world view quickly became dominant in Europe. After the publication of Newton’s *Philosophiae Naturalis Principia Mathematica*, the two scientific outlooks competed until well into the eighteenth century. Significantly—and this is reflected in the titles—Newton made mathematics indispensable for understanding the universe.

Queen Christiana of Sweden, ambitious patron of the arts and collector of learned men for her court, had seen the works of Descartes and pleaded with him to join her and teach her philosophy. She sent a man-of-war to fetch him but he was loath to go, in his words, to the “land of bears between rock and ice.” But go he did. Being more of an athlete than a scholar, the 23-year-old Queen wanted her lessons at five in the morning in a cold library with windows thrown wide open. This harsh land, where “men’s thoughts freeze during the winter months,” was too much for Descartes. A few months later he caught pneumonia and died on 11 February 1650.

Contents of the *Geometry*. The *Geometry* of Descartes is available to us in two English editions, the well-known Smith-Latham translation [3] and the only complete English translation of the whole *Discours de la Méthode* by Olscamp [4]. The latter should be consulted since the appendix on *Optics* contains much interesting material on the conics.

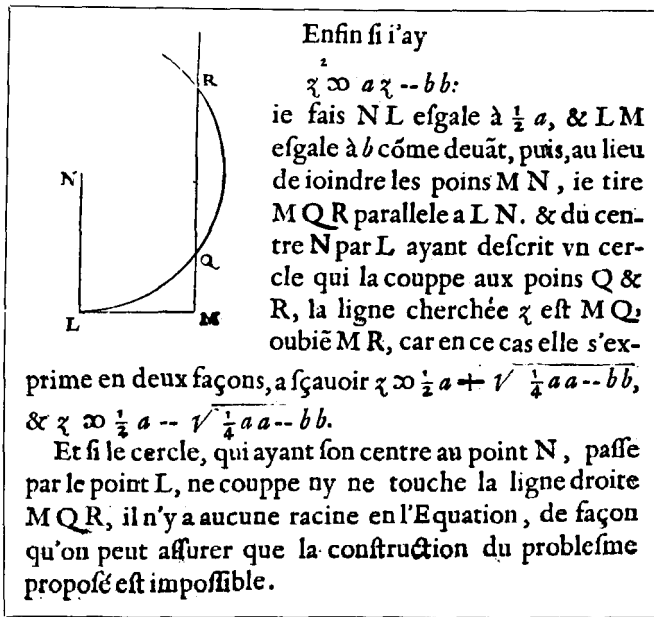


Figure 4. From p. 303 of Descartes' *Géométrie*

In the first book of the *Geometry*, Descartes gave new geometric solutions of quadratic equations. For example [Figure 4], to solve the equation $z^2 = az - b^2$ (where a and b are both positive), Descartes drew the base line LM of length b and a perpendicular line LN of length $a/2$. Then he drew the circle with center N and radius NL . This circle cut the line perpendicular to LM at M in two points. The line segments MR and MQ are the solutions of the equation, as the reader can easily check. Descartes was aware that if the circle misses (only touches) the perpendicular to LM at M , then there is no (only one) solution to the equation.

Observe [Figure 4] that we have adopted Descartes' notation. In fact, his *Geometry* is the oldest mathematics text that we can read without having great difficulties with the algebraic notation. Descartes introduced the use of x, y, z for variables and a, b, c for constants, and he also introduced the exponential notation (except that he sometimes writes " aa " for our " a^2 "). The only significant difference is that Descartes uses the symbol \propto for equality.

Another problem Descartes dealt with in the first book was the problem of Pappus (*fl.* A.D. 300–350), which he mistakenly believed was still open. The problem asks for the locus of points such that the product of the distances (measured at fixed angles) to half of a fixed set of lines is equal to the product

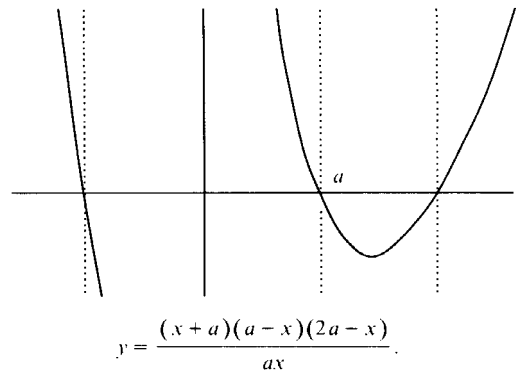


Figure 5. Cartesian Parabola

of the distances to the other half (times a constant if the number of lines is odd). If there are three or four lines, Descartes showed that the locus is a conic. As an example with five lines, Descartes considered one horizontal line and four equally spaced vertical lines [Figure 5].

He set the product of the distances to the first, third, and fourth vertical lines equal to the product of the constant distance a between the lines, the distances to the second vertical line and the horizontal line, and obtained the equation $axy = (x+a)(a-x)(2a-x)$. Newton later called this curve the Cartesian Parabola. Since there were very few curves in Descartes' day, each received its own

fancy name. This curve was only the second cubic (that is, a polynomial in two variables of degree three) ever discussed. The first was the Cissoid of Diocles (*fl. ca.* 190 B.C.). Descartes used his new curve extensively in his third book to solve equations of the fifth and sixth degrees as intersections of it and a circle.

Geometrical vs. Mechanical Curves. The second book of Descartes’ *Geometry* begins with a discussion of those curves which Descartes believed should be admitted into geometry. He does not consider the equation to be a sufficient representation of a curve, for equations are clearly algebraic objects. This forced him to always define curves by giving some geometric criterion. Later he derived the equation.

Descartes made a strict distinction between the curves that he called “geometrical” and those which he called “mechanical”, but his explanation was none too clear. It has turned out that Descartes’ geometrical (mechanical) curves are just the graphs of our algebraic (transcendental) functions. See Bos [1] for a full discussion. Descartes said that a curve is geometrical if it “can be conceived of as described by a continuous motion” [3, p. 43]. This excludes the spiral and the quadratrix because “they must be conceived of as described by two separate movements whose relation does not admit of exact determination” [3, p. 44]. Descartes allowed the use of a loop of thread to trace out a geometrical curve, as long as the shape of the string remained polygonal [3, p. 91]. Thus, the ellipse is a geometrical curve since it can be traced out using the familiar gardener’s construction using string and pegs. In *La Dioptrique*, Descartes showed how to construct the hyperbola using straightedge and string [4, p. 135]. However, the curve generated by the moving end of a piece of thread as it unwinds from a spool is a mechanical curve, for the thread was curved while wound around the spool and straight after it unwinds.

On the other hand, geometry should not include lines that are like strings, in that they are sometimes straight and sometimes curved, since the ratios between straight and curved lines are not known, and I believe cannot be discovered by human minds, and therefore no conclusion based upon such ratios can be accepted as rigorous and exact. [3, p. 91]

That straight and curved lines cannot be compared is an old dictum of Aristotle. Descartes’ adoption of

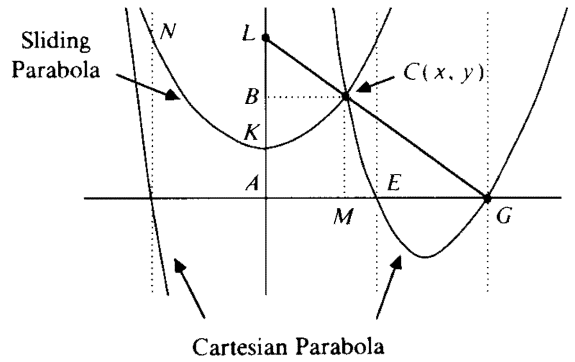


Figure 6. The Cartesian Parabola is geometric

it was important for it set up the question of rectification of curves—that is, the problem of finding arc length of curves.

Let us now consider Descartes’ argument for the Cartesian Parabola being a geometrical curve. He gave the following definition of a geometrical curve, then found its equation. Since its equation is the same as that of the Cartesian Parabola, the Cartesian Parabola is a geometrical curve.

I shall consider next the curve CEG [Figure 6], which I imagine to be described by the intersection of the parabola CKN (which is made to move so that its axis KL always lies along the straight line AB) with the ruler GL (which rotates about the point G in such a way that it constantly lies in the plane of the parabola and passes through the point L). [3, p. 84]

If we let AB be the y -axis and AG be the x -axis (Descartes used the opposite convention), then the Cartesian Parabola is the locus of all points $C(x, y)$ of intersection of the parabola that slides up and down the y -axis and the ruler that pivots at the fixed point $G(2a, 0)$ and passes through the point L moving along the y -axis with the parabola. The parabola has equation $x^2 = az$, where $a = KL$ and $z = BK$ (the focus of the parabola is one-fourth of the way from K to L). Descartes found the equation of the curve using classical geometry: Since the triangles GMC and CBL are similar, $GM/MC = CB/BL$, that is, $(2a - x)/y = x/BL$. Thus, we have

$$BK = a - BL = a - \frac{xy}{2a - x}.$$

But the equation of the parabola CKN can be written $BK = x^2/a$. Equating these expressions for BK , and simplifying, we obtain,

$$x^3 - 2ax^2 - a^2x + 2a^3 = axy,$$

which is the equation of the Cartesian Parabola. (Note that the name comes from the fact that a parabola is sliding up and down the line.)

Descartes' Subnormal Method. In our calculus classes, one important problem is to find an equation of the tangent line to a curve at a given point on the curve. Problems were not phrased this way in the seventeenth century, because equations of lines was not a well-developed topic. They asked (equivalently) for the subnormal for a given point on the curve, that is, the length of the segment on the x -axis between the abscissa of a point on the curve and the x -intercept of the normal line at that point. The subtangent was defined analogously.

Descartes presented a method for finding the subnormal [Figure 7]. If we can find a circle, with center P on the x -axis, that cuts a curve in precisely one point C , then the radius at that point is normal to the curve. But if the center of the circle through the point C be moved "ever so little" along the x -axis, the circle will cut the curve at two points. This idea provided a means of finding the subnormal for any point (x_0, y_0) on the curve. Starting with the equation of the curve and the equation of a variable circle with center $P = (v, 0)$, find the equation giving their intersection. Then choose P so that the intersection equation has a double root.

Let us consider the case of the parabola $y^2 = kx$. The circle having center $P = (v, 0)$ and radius s that passes through the point (x_0, y_0) has equation

$$(v - x_0)^2 + y_0^2 = s^2.$$

Since (x_0, y_0) is on the parabola, $y_0^2 = kx_0$, and we obtain

$$x_0^2 + (k - 2v)x_0 + (v^2 - s^2) = 0.$$

This equation will have a double root if and only if the discriminant is zero; in which case,

$$x_0 = -(k - 2v)/2, \quad \text{or} \quad vx_0 = k/2.$$

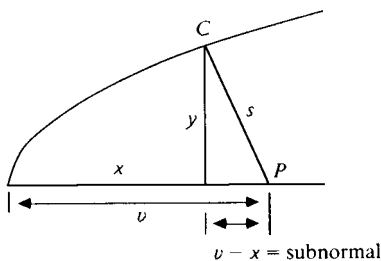


Figure 7. Finding the subnormal

This looks mysterious today, but any mathematically literate contemporary of Descartes would know that the parabola has constant subnormal. Perhaps we should check this result using the new calculus: If $y^2 = kx$, then $2yy' = k$. So $y' = k/(2y)$. Thus, the normal line at (x_0, y_0) has slope $-y_0/(k/2)$. To plot the normal line, we go down y_0 from the point (x_0, y_0) to land on the x -axis, and then go right the constant distance $k/2$. Thus, the subnormal for any point on this parabola does indeed have constant length, $k/2$. Descartes was justly proud of this work, for he wrote:

I have given a general method of drawing a straight line making right angles with a curve at an arbitrarily chosen point upon it. And I dare say that this is not only the most useful and most general problem in geometry that I know, but even that I have ever desired to know. [3, p. 95].

There is one final quotation from Descartes that is important here, for it deceived Newton in a positive way:

When the relation between all points of a curve and all points of a straight line is known [that is, when we have the equation of the curve] . . . it is easy to find . . . its diameters, axes, center and other lines [e.g., tangent and normal lines] or points which have especial significance for this curve . . . By this method alone it is then possible to find out all that can be determined about the magnitude of their areas, and there is no need for further explanation from me. [3, p. 92]

Newton believed Descartes' claim, that from the equation of a curve one can tell everything about it. This encouraged Newton to develop the variety of *ad hoc* techniques which he learned from the works of Descartes and Wallis into algorithms for solving problems about all curves. This was just one of the motivations that Newton had for inventing the calculus.

For further information about Descartes, see the DSB article by Crombie, Mahoney and Brown [6, IV, pp. 51–65]. The book by Scott [23] contains a detailed discussion of his mathematical work. Bos [1] gives an interesting study of Descartes' concept of curve. Of course, one should read the *Geometry* itself [3], [4].

Frans van Schooten (1615–1660). Schooten enrolled at the University of Leiden at age 16, where

he was carefully trained by his father in the Dutch school of algebra. He met Descartes when the latter was in Leiden to supervise the printing of the *Discours de la Méthode* (1637). Schooten recognized the value of the work but had difficulty mastering its contents. So he went to Paris for further study, where he was cordially welcomed by Mersenne.

While in Paris, Schooten read the manuscripts of Pierre de Fermat (1601-1665) and Françoise Viète (1540–1603), and under commission of the famous Leiden publishing house of Elsevier, gathered all the printed works of Viète. This included Viète's most famous work, *In Artem Analyticam Isagoge* (*Introduction to the Analytic Art*) of 1591, which dealt mainly with the theory of equations. Because of this work, Viète is known as the father of algebra. Conscious of the great importance of the scattered works of Viète on algebra, geometry, and analysis, which had been published separately from 1579 to 1615, Schooten republished them with commentary as *Francisci Vietae Opera Mathematica* (1646). The work quickly became an indispensable collection of mathematical source materials, and Newton carefully studied a copy from the Cambridge libraries [20, I, p. 21].

Schooten returned to Leiden in 1643 and began working on a Latin translation of Descartes' *Géométrie*, which he published in 1649. Descartes had been dissatisfied with the form and argument of his *Géométrie* from the very day of its publication, and therefore encouraged the writing of commentaries clarifying its obscurities and developing its approach. Because of its valuable commentary and excellent figures, Schooten's edition was enthusiastically received. This success led him to prepare a much enlarged second edition that appeared in two volumes (1659–1661). It contained about 800 pages of commentary and new work, in addition to the 100 page translation of Descartes' *Géométrie*, and included [20, I, pp. 19–20]:

- Schooten's extremely valuable commentaries. Many of these details were derived directly from Descartes' own criticisms made in correspondence with Schooten.
- Florian Debeaune's (1601–1652) *Notae Breves*, a work which Descartes welcomed as a perceptive exposition of the more elementary aspects of his work. Debeaune posed the first inverse tangent problem.
- Jan Hudde's (1628–1704) studies on equations and extreme values. His rule for locating dou-

ble roots of equations was useful in applying Descartes' tangent method. It was an important precursor of the derivative.

- Jan de Witt's (1629–1695) excellent tract on conic sections.
- An example of Fermat's extreme value and tangent method.
- Christiaan Huygens' (1629–1695) first publication, an improved method for finding the tangent to the conchoid.
- Hendrik van Heuraet's (1633–ca. 1660) rectification method, of which we shall say more below.

All of this shows the great effort that Schooten devoted to the training of his students and to the dissemination of their findings. Much of their work is available only in correspondence, careful studies of which are currently being made. It was from these editions of Schooten that mathematicians learned of the work of Descartes. It was the second Latin edition that Newton borrowed and annotated in the summer and autumn of 1664 (the copy he bought the following winter may have been the 1649 edition). It had an immense impact on his mathematical development; for after mastering it, he was current with research in the new analysis.

John Wallis (1616–1703). Before attending Emmanuel College, Cambridge, the only mathematics Wallis knew was what he learned from his brother who was preparing for a trade. At Cambridge, mathematics "were scarce looked upon." He took his M.A. in 1640 and was ordained. In 1649, he was appointed Savilian Professor of Geometry at Oxford, an appointment that must have surprised those who thought the only mathematics he had done was to decode a few messages for the Parliamentarians.

This is not quite true, but, wrote Wallis "I had not then [in 1648] seen Descartes' *Geometry*." [20, III, p. xv]. In 1647 or 1648, he chanced upon Oughtred's *Clavis*, mastering it in a few weeks, and then rediscovered Cardano's formula for the cubic. In 1648, at the request of Cambridge professor of mathematics John Smith, he reworked Descartes' treatment of the fourth-degree equation by factoring it into two quadratics. As soon as he was appointed Savilian Professor at Oxford, he took up the study of mathematics, with rare energy and perseverance, and soon became one of the best mathematicians in Europe. He held the post for 50 years.

Wallis's *Operum Mathematicorum Pars Altera* (Oxford, 1656) was a fat and rather motley two-part collection of his early mathematical lectures, commentaries, and researches [20, I, p. 23]. It contained his *De Sectionibus Conicis* (dated 1655), a treatise of 110 pages that was the first elementary text on the conics treated from the Cartesian viewpoint. In an appendix, Wallis tried to extend the approach to higher plane curves, especially the cubical parabola $a^2y = x^3$, where the constant a^2 was used to preserve dimensionality. He successfully found the sub-tangent, but had trouble with the graph because he did not feel comfortable with negative numbers. He also introduced the semi-cubical parabola $ay^2 = x^3$, a curve that played a very important role in the development of the calculus [30, pp. 295–298]. Quite suddenly the mathematical world had been presented with a powerful analytic geometry, only to find that there were few curves on which to practice it. The new perspective of Wallis—which took some time to be adopted by the mathematical community—was that any algebraic equation in two variables defines a curve [13, p. 238].

Together with his conic sections, Wallis published the work on which his fame rests *Arithmetica Infinitorum* (dated 1656; printed 1655). This volume developed from his study of the *Opera Geometrica* (1644) of Torricelli (1608–1647). Wallis tried to apply these methods to the quadrature of the circle, but not even the study of the voluminous *Opus Geometricum* (1647) of Gregorius Saint Vincent (1584–1667), helped. Out of the project of squaring the circle, he did get his famous infinite product for $\pi/4$.

The *Arithmetica Infinitorum* exerted a singularly important influence on Newton when he studied it in the winter of 1664–1665. From it, Newton learned of the problem of quadratures, or, as we now say, finding areas under curves. Newton probably also read Wallis's *Tractatus Duo* (1659) that presented his research on the cycloid, cissoid, and other geometrical figures.

Rectification of Curves. By 1638, Descartes suspected that the logarithmic spiral might be rectifiable; that is, the length of an arc of the curve could be computed. Even if correct, this would not cause him any difficulties because the spiral is a mechanical curve, and Descartes only accepted Aristotle's dictum that straight lines and curved lines could not be compared for geometrical curves. In 1657, Huygens found the length of an arc of a parabola; but he used a mechanical curve in his solution, and thus

Descartes' version of Aristotle's dictum was still intact. Also Huygens' method did not generalize.

The first geometrical curve to be rectified in a geometric way was Wallis's semi-cubical parabola $ay^2 = x^3$. As often happens, several people solved the problem simultaneously: William Neil in 1657, Hendrick van Heuraet in 1659 [14], and Pierre de Fermat in 1660. Of course, a priority dispute erupted. Heuraet's solution was the most influential because it was published in Schooten's second Latin edition of Descartes' *Geometry*. The proof used the new classical geometry of the seventeenth century and was fairly intricate (for details, see [8] or [13]). The method of proof was to replace the problem of rectification of the semi-cubical parabola by a simpler problem, the quadrature of an ordinary parabola.

This transformation of the problem to a simpler one shows up even when we do the problem today with the calculus, but it is so slick that it is easy to miss what happens. Starting with $y^2 = x^3$ (it is no accident that we still do this first today), we obtain $(y')^2 = 9x/4$. Thus, the arc length from, say, $(0, 0)$ to $(4, 8)$, is

$$L = \int_0^4 \sqrt{1 + (9x/4)} dx.$$

The substitution $u = 1 + (9x/4)$ transforms this into

$$L = (4/9) \int_1^{10} \sqrt{u} du.$$

The first of these integrals represents an arc length, whereas the second stands for the area under a parabola. Today, we just look at these as two simple integration problems, but in the old days B(efore) C(alculus), these were viewed as two separate kinds of problems.

Heuraet's method was entirely general. When Newton saw the proof, he realized the value of transforming one type of problem into another. This is one of the roots of the Fundamental Theorem of Calculus. It is the biggest swap of all—we trade integration for anti-differentiation. This is precisely what Newton did soon after he read Heuraet's proof. (For a full history of the rectification problem, see Hofmann [11, Ch. 8].)

Concluding Remarks about Newton's Readings.

In order to do creative work, a mathematician “needs an adequate notation, a competent knowledge of mathematical structure and the nature of axiomatic proof, an excellent grasp of the hard core of existing mathematics and some sense of promising line for

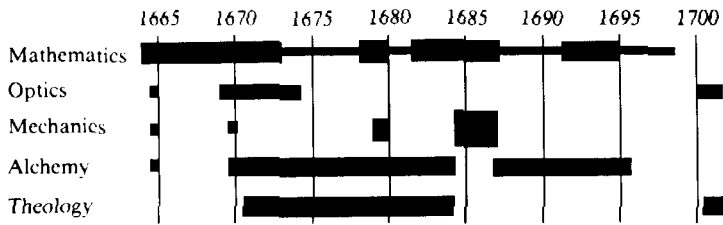


Figure 8. Newton’s areas of activity

future advance.” [20, I, p. 11]. The works that Newton chose to read in 1664 and 1665 magnificently met these needs. He took his arithmetic symbolism from Oughtred, his geometrical form from Descartes. Of course, he grafted on new modifications of his own while creating the calculus. He learned elementary scholastic logic in grammar school and traditional forms of mathematical proof from Euclid. He learned the new analytic geometry of the seventeenth century from Schooten and de Witt, topics in algebra and the theory of equations from Viète, Oughtred, Schooten and Wallis. Most importantly, he learned of the twin problems of infinitesimal analysis: From Descartes, the method of tangents; from Wallis, quadrature. There were plenty of open problems for Newton to attack. Without doubt, the two strongest influences on Newton were Descartes and Wallis. [20, I, pp. 11–13]

It is of as much interest to note *what Newton did not read*. We miss the names of Napier, Briggs, Harriot, Desargues, Pascal, Fermat, Stevin, Kepler, Cavalieri, and Torricelli. Among the Greeks there is only Euclid, not Apollonius nor Archimedes. In fact, Newton seemed to dislike the method of exhaustion. There is great significance in this lack of knowledge of ancient mathematics and of the new classical (as opposed to analytic) geometry of the seventeenth century. He was not hampered by its knowledge. Had Newton gained a deep knowledge of classical geometry and the new classical geometry of his century, I conjecture it would have hindered his invention of the calculus (and similarly for Leibniz who was also ignorant of classical geometry).

As Westfall points out [28, p. 100] about Newton’s readings:

In roughly a year, without benefit of instruction, he mastered the entire achievement of seventeenth-century analysis and began to break new ground.

In fact by mid-1665 Newton’s urge to learn from others seems to have abated [20, I, p. 15].

3 Newton’s works

Newton was an extraordinary scientist because he made so many fundamental contributions to different fields:

- Mathematics, both pure and applied
- Optics and the theory of light and color
- Design of scientific instruments
- Synthesis and codification of dynamics
- Invention of the concept and law of universal gravity.

In addition, we now know, and are willing to admit, that he spent immense amounts of time working on:

- Alchemy
- Chronology, church history, and interpretation of the Scriptures.

The range and depth of Newton’s intellectual pursuits never ceases to amaze us.

As a first step in understanding Newton’s contributions, consider the chart above that indicates when Newton was involved in various research areas. One might think that Newton thought about everything all of the time, but the manuscript record shows that he worked on only a few areas at any one time, and these were not necessarily — in his mind at least — disjoint.

We begin with a synopsis of Newton’s mathematics as presented in Whiteside’s edition of Newton’s *Papers* [20]. This will be followed by a thumbnail sketch of each of these areas of Newton’s intellectual efforts. Since it is impossible to discuss all of his contributions here, only a few examples of Newton’s mathematical work will be discussed in detail. These were chosen with the teacher in mind, to provide examples that can be used in the classroom.

Volume I. (1664–1666). The volume begins with Newton’s annotations on the works of Oughtred, Descartes, Schooten, Viète, and Wallis. The bulk

consists of research on analytic geometry and the calculus. Newton turns Descartes' subnormal technique into the notion of curvature, and Hudde's rule for double roots into fluxions (differentiation). We see the calculus become an algorithm in mid-1665. This early work on the calculus was summarized in the October 1666 tract on fluxions. In a schematic diagram, Whiteside [20, I, p. 154] shows how all of these ideas came together to give birth to the calculus. The volume ends with miscellaneous work on trigonometry, the theory of equations, and geometrical optics.

Volume II. (1667–1670). Work on classification of cubics begins here and was published as an appendix to his *Optics* (1704). In this volume, we see Newton struggling with the graphs. The most important work on the calculus is the hastily composed 1669 tract *De Analysisi* that summarizes all of his work thus far. He gave a copy of this to Barrow in 1669 to assert his priority over Nicolaus Mercator (1619–1687) whose *Logarithmotechnia* (1668), with its infinite series for the logarithm, had just appeared. Half the volume consists of his annotations on the *Algebra* of Kinckhuysen. One piece of Newton's advice here is too good not to pass on to our students:

After the novice has exercised himself some little while in algebraic computation . . . I judge it not unfitting that he test his intellectual powers in reducing easier problems to an equation, even though perhaps he may not yet have attained their resolution. Indeed, when he is moderately well versed in this subject . . . then will he with greater profit and enjoyment contemplate the nature and properties of equations and learn their algebraic, geometrical and arithmetical resolutions. [pp. 423–425]

Volume III. (1670–1673). Although Barrow encouraged Newton to revise *De Analysisi* for publication, the booksellers were uninterested. But he did combine the two earlier works on the calculus and many new results in a 1671 tract, with an important foundational change: he postulated a fluent variable of time for his fluxions, that is, all his derivatives are time derivatives. Also, here is an investigation of Huygens' pendulum clock and more research on geometric optics.

Volume IV. (1674–1684). Research in theology and alchemy kept him busy (Figure 8), though his work on mathematics never entirely stopped. This volume contains some of Newton's research on algebra,

number theory, trigonometry, and analytical geometry. In the middle of this period, he became fascinated with the classical geometry of the Greeks. Only at the end of this period did Newton show great interest in fluxions and infinite series.

Volume V. (1683–1684). The bulk of this volume consists of Newton's ninety-seven self-styled "lectures", deposited as his Lucasian lectures on algebra for the period 1673–1683. The *Arithmetica Universalis* given here is an incomplete revision of the algebra lectures. Its published version was his most read work, not the papers on calculus.

Volume VI. (1684–1691). Halley's visit in August 1684 turned Newton's interest to the geometry and dynamics of motion, the subject of this entire volume. The work dates from the period 1684–1686, and is arguably as creative as the miracle years of 1664–1666.

Volume VII. (1691–1695). In the early winter of 1691–1692, Newton wrote *De Quadratura Curvarum*, on the quadrature of curves. He also dealt with classical geometry (1693), higher plane curves, and finite-difference approximations (1695). As always, Whiteside has "taken care to preserve all the significant idiosyncrasies, contractions, superscripts and archaic spellings" of the "ink-blobbed, much-cancelled and often rudely scrawled manuscripts." [p. ix]

Volume VIII. (1697–1722). Most mathematicians will find this the most interesting volume after the first, for it contains Newton's solution (simply stated without proof) of the brachistochrone problem as well as documents related to the priority dispute. (To see that this dispute involved much more than mathematics, read Hall's *Philosophers at War* [9].)

We calculus teachers should refrain from telling our students that Newton invented the calculus because he was motivated by physical considerations. Although applications are an excellent reason for studying the calculus, in Newton's case the record is clear: first mathematics, then applications.

The Binomial Theorem. On the frontispiece of the first volume of Newton's Papers we see the manuscript where he took up the age old problem of squaring the circle, or (to make the activity sound more respectable) the quadrature of the circle. He became interested in this problem after reading Wallis's *Arithmetica Infinitorum*. Newton learned there how to evaluate the integrals (here expressed in Leibniz's notation) $\int_0^x (1-x^2)^{n/2} dx$, where n is an

$n =$	-2	0	2	4	6	8	...	times
	1	1	1	1	1	1	...	x
	-1	0	1	2	3	4	...	$-x^3/3$
	1	0	0	1	3	6	...	$x^5/5$
	-1	0	0	0	1	4	...	$-x^7/7$
	1	0	0	0	0	1	...	$x^9/9$
	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Table 1.

$n =$	-3	-2	-1	0	1	2	3	4	5	6	7	8	...	times
	1	1	1	1	1	1	1	1	1	1	1	1	...	x
	$-\frac{3}{2}$	-1	$-\frac{1}{2}$	0	$-\frac{1}{2}$	1	$\frac{3}{2}$	2	$\frac{5}{2}$	3	$\frac{7}{2}$	4	...	$-x^3/3$
	$\frac{15}{8}$	1	$\frac{3}{8}$	0	$-\frac{1}{8}$	0	$\frac{3}{8}$	1	$\frac{15}{8}$	3	$\frac{35}{8}$	6	...	$x^5/5$
	$-\frac{35}{16}$	-1	$\frac{5}{16}$	0	$\frac{1}{16}$	0	$-\frac{1}{16}$	0	$\frac{5}{16}$	1	$\frac{35}{16}$	4	...	$-x^7/7$
	$\frac{315}{128}$	1	$\frac{35}{128}$	0	$-\frac{5}{128}$	0	$\frac{3}{128}$	0	$-\frac{5}{128}$	0	$\frac{35}{128}$	1	...	$x^9/9$
	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Table 2.

even integer. Newton tabulated the values of these integrals in his attempt to find the area of a circle ($n = 1$). To see how he did this consider the case when $n = 6$:

$$\int_0^x (1 - x^2)^{6/2} dx = 1(x) + 3(-x^3/3) + 3(x^5/5) + 1(-x^7/7).$$

The factors in parentheses are recorded in the rightmost column of the table below. The coefficients, 1, 3, 3, 1, are recorded in the column labeled $n = 6$. In general, to evaluate $\int_0^x (1 - x^2)^{n/2} dx$, sum the products of the values in the n th-column by the corresponding terms in the rightmost column.

$n =$	0	2	4	6	8	...	times
	1	1	1	1	1	...	x
		1	2	3	4	...	$-x^3/3$
			1	3	6	...	$x^5/5$
				1	4	...	$-x^7/7$
					1	...	$x^9/9$
					⋮	⋮	⋮

Wallis had also tabulated these integrals, but since he used 1 rather than x as an upper limit, he did not see the pattern. But Newton recognized it as "Oughtreds Analytical table", from his readings of Oughtred's *Clavis* [20, I, p. 452]. We, of course, now call this Pascal's triangle. Newton knew that each number in the table is the sum of the number

to its left and the one above that, so he decided to extend the pattern backwards for all even values of n . Thus he obtained Table 1.

To extend this table to odd values of n , Newton used a complicated proportionality argument (see [31] for details). Later, in a letter to Leibniz [19, I, pp. 130-131], Newton provided an easier explanation for the extension. When n is even, say, $n = 2m$, the k th entry in the n th column is given by the binomial coefficient

$$\binom{m}{k-1} = m!/k!(m-k)!$$

Newton ignored the restriction that n must be even and used the formula for binomial coefficients when n was odd. For example, the fourth entry in the $n = 1$ column is given by

$$\binom{\frac{1}{2}}{4-1} = \frac{(\frac{1}{2})(\frac{1}{2}-1)(\frac{1}{2}-2)}{(1)(2)(3)}.$$

Thus, he obtained Table 2.

Now, from the $n = 1$ column, Newton was able to draw the conclusion that he sought:

$$\int_0^x (1 - x^2)^{1/2} dx = x + \frac{1}{2} \left(-\frac{x^3}{3} \right) + -\frac{1}{8} \left(\frac{x^5}{5} \right) + \frac{1}{16} \left(-\frac{x^7}{7} \right) + \dots$$

$$\sqrt[m]{P + PQ} = P^{\frac{m}{n}} + \frac{m}{n} A Q + \frac{m-2}{2n} B Q^2 + \frac{m-3}{3n} C Q^3 + \dots$$

Where $P + PQ$ is the Quantity, whose Root is to be extracted, or any Power formed from it, or the Root of any such Power extracted. P is the first Term of such Quantity; Q , the rest (of such proposed Quantity) divided by that first Term, And $\frac{m}{n}$ the Exponent of such Root or Dimension sought. That is, in the present case, (for a Quadratick Root,) $\frac{1}{2}$.

Figure 9. First publication of the Binomial Theorem, 1685

For $x = 1$, this gives an infinite series for the area of (a quadrant of) a circle. From this, Newton jumped to the conclusion that a similar “interpolation” could be done on curves (we would say, on functions) as well as on their quadratures (integrals), and then guessed the Binomial Theorem for fractional exponents. He checked this result several ways. First, he formally used the square root algorithm to obtain the series

$$(1 - x^2)^{1/2} = 1 - \frac{1}{2}x^2 - \frac{1}{8}x^4 - \frac{1}{16}x^6 - \dots$$

Then he checked that it agreed with the Binomial Theorem. Next, he squared both sides of the above equation to see that an equality resulted. As a further check, he used formal long division to obtain an infinite series for $(1 + x)^{-1}$. Note the wonderful research techniques he is using. Nonetheless,

The paradox remains that such Wallisian interpolation procedures, however plausible, are in no way a proof, and that a central tenet of Newton’s mathematical method lacked any sort of rigorous justification . . . Of course, the binomial theorem worked marvellously, and that was enough for the 17th century mathematician. [31, p. 180]

Newton became tremendously excited with his new tool, the Binomial Theorem, which became a mainstay of his newly developing calculus. He also did such bizarre computations as approximating $\log(1.2)$ to 57 decimal places.

The Binomial Theorem was Newton’s first mathematical publication. It appeared in Wallis’s *Treatise of Algebra* (Figure 9) in a summary of Newton’s two famous letters to Leibniz in 1676 [24, pp. 330–331]. These letters are readily available, with ample commentary, in Newton’s *Correspondence* [19, II, pp. 20–47 and 110–161].

Optics. Newton’s earliest work on optics was done at Cambridge and the experiments continued at Woolsthorpe during the plague, but was not put in near final form until he was preparing his Lucasian lectures for 1670–1672. It had long been known (see, for example, Descartes [4, p. 335]) that when light passed through a prism it was dispersed into a colorful spectrum. Newton was able to give a quantitative analysis of this behavior and to devise a new theory of light. In February 1671/72 (the slash date was used because England had not yet adopted the Gregorian calendar), this resulted in Newton’s first publication in optics, the lengthy title of which also provides an abstract:

A Letter of Mr. Isaac Newton, Mathematick Professor in the University of Cambridge; containing his New Theory about Light and Colors: Where Light is declared to be not Similar or Homogeneal, but consisting of difform rays, some of which are more refrangible than others: And Colors are affirm’d to be not Qualifications of Light, deriv’d from Refractions of natural Bodies, (as ’tis generally believed;) but Original and Connate properties, which in divers rays are divers: Where several Observations and Experiments are alleged to prove the said Theory. [18, p. 47]

This work engendered a controversy with Robert Hooke (1635–1703), who claimed to have published the ideas earlier. As a consequence, Newton became extremely reluctant to publish. In fact, the *Optics* was not published until 1704, the year after Hooke’s death.

In developing his theory of light, Newton realized that lenses caused chromatic aberration. This set him thinking about telescope design, and he concluded that the problem could be avoided by using mirrors instead of lenses. Consequently he designed

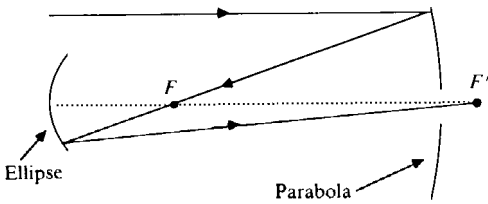


Figure 10. Gregorian Telescope, 1663.

a reflecting telescope, built one himself, and then described it in the March 25, 1672 issue of the *Philosophical Transactions*. These first papers of Newton have been photoreproduced by I. B. Cohen [18], along with a valuable introduction by Thomas Kuhn. Rather than describe Newton's theory of light (which has been done by Alan Shapiro in the first volume of *The Optical Papers of Isaac Newton* [21]), we shall briefly discuss telescope design. This provides an interesting classroom example of the reflective properties of the conics.

The first reflective telescope was designed by James Gregory (1638–1675) and published in his *Optica Promota* of 1663, a work which Newton did not read until after he had invented his own telescope. Gregory's telescope consists of a concave primary mirror (on the right in Figure 10) that is parabolic in shape, and a concave secondary mirror that is elliptical (strictly speaking, the surfaces generated by rotating these conics about the axis of the telescope). The incoming rays of starlight bounce off the parabolic mirror and are reflected through its focus. Beyond that focus is an elliptical mirror that shares a focus with the parabola and has its other focus behind a small hole in the primary mirror. Thus, after the reflected rays of starlight pass through the common focus of the parabola and ellipse, they are reflected off the elliptical secondary mirror and converge at the second focus of the ellipse. Gregory tried to have a telescope built to his design, but the opticians were unable to polish the mirrors properly.

In 1668, Newton placed a flat secondary mirror between the primary parabolic mirror and its focus

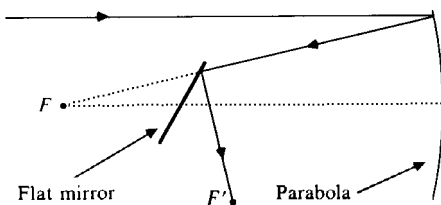


Figure 11. Newtonian Telescope, 1668.

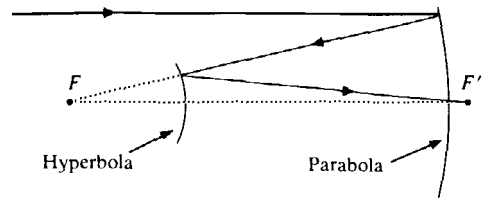


Figure 12. Cassegrain Telescope, 1672

[Figure 11]. The eyepiece was located at the side of the telescope. Incoming rays of starlight reflect off the parabolic mirror and head for its focus F . Before they get there, they are reflected off the flat mirror. Then they converge toward F' , the point symmetric to F with respect to the plane of the flat mirror. This invention remained unknown until Newton made another one (casting and polishing the mirrors himself) and presented it to the Royal Society of London on 11 January 1672. This so impressed the members that they elected him a Fellow of the Royal Society at that very same meeting.

Later in 1672, another telescope design [Figure 12] was published by Guillaume Cassegrain (fl. ca. 1672) in France and abstracted in the *Transactions of the Royal Society*. The concave primary mirror is again a parabola with a hole in the center, and the secondary is a convex hyperbolic mirror which shares a focus with the parabola and has its other focus behind the hole in the parabola. Rays of starlight reflected from the primary parabolic mirror head toward the focus of the parabola. Before reaching that focus, they are reflected by the hyperbolic mirror toward the other focus of the hyperbola.

Cassegrain claimed that his design was superior to Newton's. In what was to become his typical style, Newton marshalled his evidence and attacked furiously. He claimed that Cassegrain's idea was not only a minor modification of Gregory's, but also optically inferior. Cassegrain retreated into anonymity. But Newton was wrong about the superiority of the Gregorian telescope. In 1779, Jesse Ramsden (1735–1800) showed that the combination of a concave and a convex mirror partially corrects the spherical aberrations, whereas in the Gregorian telescope, the aberrations of the two concave mirrors are additive. Today the Cassegrain model is used in most large reflectors.

A mathematical result of Newton's work on optics grew out of the problem of grinding a hyperbolic mirror (although he did not use one, the possibility of a hyperbolic lens was noted by Descartes [4, p. 139]). Newton, and independently Christopher Wren

(1632–1723), discovered that the hyperboloid of one sheet was a ruled surface. Newton used this result to show how to make a hyperboloid of one sheet on a lathe by holding the chisel obliquely to the axis of the lathe.

Religion. Despite inheriting his stepfather's theological library and buying several theological books when he came to Cambridge, Newton's serious study of theology began only in the early 1670's. No doubt this came about because the position of Fellow at Trinity required that one had to be ordained in the Anglican Church within seven years of receiving the M.A. In Newton's case, this was by 1675. Not being one to do anything halfway, Newton became engaged in an extensive reading program that took him through all the early Church Fathers. As a result, ordination became impossible for he had become a heretic.

Newton became an Arian or Unitarian — he denied the Trinity — of deep conviction and remained so for the rest of his life. His argument was: "Though Christ was the only begotten son of God, and hence never merely a man, he was not equal to God, not even after God exalted him to sit at his right side as a reward for his obedience unto death." [29, p. 130]. Newton arrived at this position through a careful analysis of Scripture. He believed that a deceitful Roman Church had manipulated the Emperor Theodosius to introduce the false doctrine of the Trinity into the Scriptures in the fourth century. The *Book of Revelation* was crucial to Newton's interpretation. He believed that the Roman Church was the "Great Apostasy" and never ceased to hate and fear it [28, p. 321].

By 1675, Newton was making plans to leave Cambridge for he knew that as a Fellow at the College of the Holy and Undivided Trinity he could not reveal his Arian views. To do so would be socially unacceptable, and he never in his life did so, except obliquely to a few people of similar persuasion. That he read the situation correctly is indicated by the dismissal of William Whiston (1667–1752), Newton's successor in the Lucasian Chair, for the uncompromising expression of Unitarian views. But just at this time, the Crown granted a special dispensation that the occupant of the Lucasian Professorship was not required to be ordained. Thus, Newton could stay at Cambridge. Newton's theological studies continued until work on the *Principia* interrupted [Figure 8]. In London, he was able to take up his theological studies again, and they continued for the rest of his

life. (For further details, see [28, pp. 309–334] or [29].)

Alchemy. Newton's interest in alchemy has long been embarrassing to some scholars, while others delight in this trace of hermeticism and dub him a mystic. But there is now no doubt that he was a serious practitioner [Figure 8]. From 1669 (when he bought his first chemicals) until 1684 (when work on the *Principia* interrupted), Newton spent long hours in the "laboratory". Newton again practiced alchemy from 1686 until 1696, but after he moved to London he never took it up again seriously [27, p. 121]. Newton did plan on adding alchemical references to the second edition of the *Principia* although he never did so. (For details on his alchemical work, see [5].) One benefit of this work was that he was able to cast the speculum for his first telescope.

In 1693, Newton suffered a nervous breakdown of uncertain duration and severity. There is no doubt that he frequently tasted his chemicals, but whether it was caused by mercury poisoning is debatable [7, pp. 88–90]. When Newton wrote to Oldenburg in 1673 that he intended "to be no further solicitous about matters of Philosophy" [19, I, p. 294], and to Hooke in 1679 that "I had for some years past been endeavouring to bend my self from Philosophy to other studies in so much y^t I have long grutched the time spent in y^e study" [19, II, p. 300], we must take him at his word. During most of the decade of the 1670s, Newton preferred theology and alchemy to physics and mathematics.

The *Principia*. In his old age, Newton liked to reminisce and he himself started the story of the falling apple. We have four independent accounts of the tale [7, p. 29–31]. Here is Conduitt's [28, p. 154]:

In the year 1666 he retired again from Cambridge . . . to his mother in Lincolnshire & whilst he was musing in a garden it came into his thought that the power of gravity (w^{ch}) brought an apple from the tree to the ground) was not limited to a certain distance from the earth but that this power must extend much farther than was usually thought. Why not as high as the moon.

So let us grant that a falling apple started Newton thinking about gravity during the plague years; even if he made up the story it is harmless. But his retelling of this event, in his 1718 letter to Des-Maizeaux (which we quoted earlier), is not harmless. Newton attempted to push back the date of his

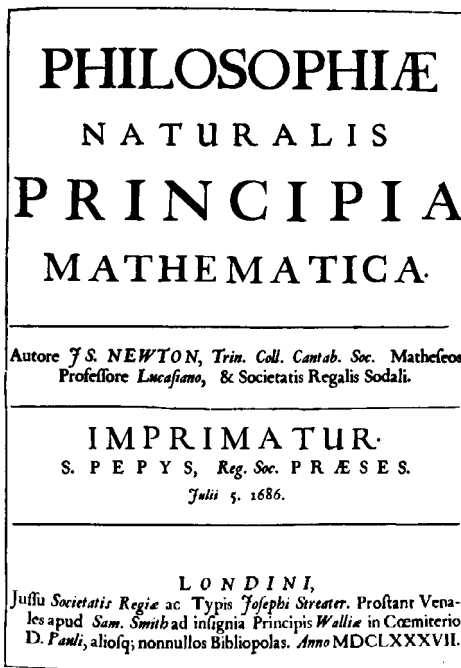


Figure 13. The *Principia*

discovery of the law of universal gravitation to the plague years. His papers tell quite a different story.

In late 1664, Newton learned Kepler’s third law: the square of the time that it takes a planet to make one elliptical revolution around the sun is proportional to the cube of the mean distance from the sun, that is, $T^2 \sim R^3$. The following January, Newton discovered the Central Force Law (see the letter to DesMaizeaux), which Huygens independently discovered and first published without proof in his *Horologium Oscillatorium* (1673) (see [12]). The Central Force Law states that the centrifugal (center fleeing) force acting on a body traveling about a central point is proportional to the square of the speed and inversely proportional to the radius of the orbit: $F = S^2/R$. Strictly speaking, this “force” is an acceleration, but we shall follow Newton’s usage.

Newton was able to discover that the gravitational force between a planet and the sun must be inversely proportional to the square of the distance between them. If a planet travels with uniform speed around a circular (not elliptical) orbit of radius R in time T , then its speed S is $2\pi R/T$. Thus,

$$F = \frac{S^2}{R} = \frac{(2\pi R/T)^2}{R} = 4\pi^2 \left(\frac{R^3}{T^2} \right) \left(\frac{1}{R^2} \right).$$

By Kepler’s third law, R^3/T^2 is constant, and hence, $F \sim 1/R^2$. Newton left off at this point, devoting

most of the next decade to alchemy and theology, though he was never completely divorced from mathematics (Figure 8).

Hooke, Halley, and Wren were able to make this same deduction by 1679, but the problem of explaining the elliptical orbits remained. On 24 November 1679, Hooke wrote to Newton suggesting a private “philosophical”, that is, scientific, correspondence on topics of mutual concern. In this letter, Hooke mentioned his hypothesis of “compounding the celestial motions of the planetts [out] of a direct motion by the tangent & an attractive motion towards the centrall body.” [19, II, p. 297]. This does not seem to be much of a hint for proving that if the inverse square law holds, then the planets must move in elliptical orbits. But it started Newton thinking about the question again. Hooke gave further encouragement on January 17, when he wrote “I doubt not but that by your excellent method you will easily find out what that Curve must be.” [19, II, p. 313]. Newton did succeed in finding the answer, but he kept it to himself.

It was also in 1679 that Newton learned of Kepler’s second law: the radius from the sun to a planet sweeps out equal areas in equal times. It seems strange that Newton would have learned of the third law as a student in 1664, but not about the second until years later. The explanation is that the third law was generally accepted in the scientific community because it could be empirically verified, whereas the second was much more of a conjecture.

In August 1684, Edmond Halley (1656–1743) visited the 41-year-old Lucasian Professor at Cambridge. He asked the question that had been consuming him and his friends Hooke and Wren at the Royal Society in London: What path would the planets describe if they were attracted to the sun with a force varying inversely as the square of the distance between them? Newton replied at once that the orbits would be ellipses. Since this was the expected answer, Halley asked Newton how he knew. Newton astonished him by answering that he had calculated it. Halley asked to see Newton’s computation, but as Newton seemingly saved every scrap of paper he ever wrote on, he (not surprisingly) could not find it. Perhaps he did not want to find it; his desire to be left alone to pursue his own interests, his fear of controversy, and his reluctance to publish would all make Newton want to carefully check his proof over again before he showed it to anyone. In November of 1684, Newton did send the computation to Halley in London, who was so excited that he prompted

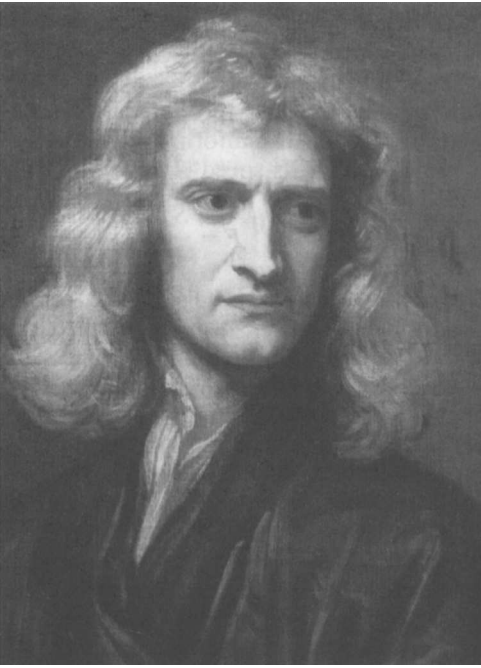


Figure 14. Isaac Newton at age 46, two years after the *Principia*'s publication

Newton to expand his work. (Weinstock [25] has challenged the common view that this proof actually appears in the *Principia*.)

Newton put aside his alchemical and theological studies to work on what was to become the most significant scientific treatise ever written: *Philosophiæ Naturalis Principia Mathematica*. It took Newton eighteen months of intense intellectual effort to compose his masterpiece, but the time was not just spent in writing up results that he had completed long ago. Many critical ideas in the *Principia* were not developed until the treatise was being written. In particular, Newton created the concept of universal gravity during this period.

Myth: *Though Newton used the notation of the calculus in arriving at his results, he was careful in the *Principia* to recast all the work in the form of classical Greek geometry understandable by other mathematicians and astronomers.*

Newton started this myth himself in the midst of the priority dispute with Leibniz. If he could argue that he had used his calculus in composing the *Principia*, then he could claim that he did not steal the calculus from Leibniz who published his first paper on the (differential) calculus in 1684.

The method of fluxions [Newton's calculus] is

intrinsically algebraic rather than geometrical, and there is not the slightest reason — in the historical evidence or in logic — to suppose that the argument of the *Principia* was ever cast in an algebraic rather than the geometric mode in which it was published. [9, p. 28]

The geometrical format of the *Principia* is explained by the fact that around 1678, Newton became fascinated with classical geometry [Figure 8]. The *Principia* appears to be densely packed classical geometry, but that is only a facade. One need only read a bit to realize that it is packed with the informal geometrical ideas of the new analysis, the calculus. However, the formal machinery of the algebraic algorithms of the calculus is not to be found there. In order to make this point clear, it would help to look at the proof of a proposition from the *Principia*. Book I, Section II, Proposition I, Theorem I says:

The areas which revolving bodies describe by radii drawn to an immovable centre of force do lie in the same immovable planes, and are proportional to the times in which they are described. [17, p. 40].

That is, if the gravitational force (whatever it might be) always acts toward a fixed point S , then Kepler's equal area law holds.

Newton's proof begins with classical geometry [Figure 15]. Suppose we consider equal time intervals, and that the body moves from A to B in one of those intervals. In the next interval it would move, on the same straight line, from B to c if no external force acts on it. The triangles SAB and SBc that are swept out in these equal time intervals have equal areas since the bases AB and Bc are equal, and the triangles have the same altitude. However, if at B "a centripetal [center seeking] force acts at once with a great impulse", then the body moves to

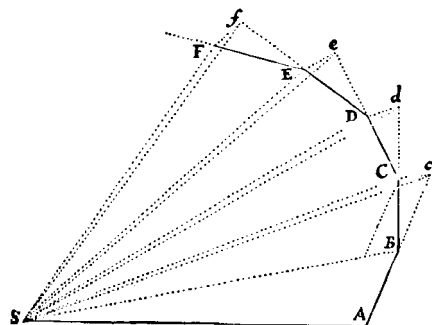


Figure 15. Newton's *Principia*, p.37

some other point C (in the same plane) in the next time interval. The Parallelogram Law of Forces determines the location of the point C ; the lines Cc and SB are parallel. Triangles SBc and SBC also have the same area, since they have a common base SB and their altitudes are equal, namely, the distance between the parallel lines SB and Cc . By transitivity, the triangles SAB and SBC have equal areas. Similarly for other triangles in the diagram. So far the proof was easy geometry. Next, Newton used an idea from his calculus: “Now let the number of those triangles be augmented, and their breadth diminished in infinitum . . . their ultimate perimeter ADF will be a curved line: and therefore the centripetal force, by which the body is continually drawn back from the tangent of this curve, will act continually” and the areas traced out in equal times will be equal [17, pp. 40–41].

So that was really quite easy. The geometric ideas of the calculus are used constantly in the *Principia*, but the algebraic notations are not.

4 Conclusion

On 5 February 1675/76, Newton wrote to Hooke [19,1, p. 416]:

What Des-Cartes did was a good step. You have added much . . . If I have seen further it is by standing on ye sholders of Giants.

While it is important to realize that Newton recognized the contributions of his predecessors, we must by now feel that Newton was the greatest giant of all. Just as Westfall, after twenty years of effort preparing *Never at Rest*, was more in awe of Newton when he finished than when he began, we too may realize that the closer we get to Newton, even when standing on the shoulders of Whiteside, the bigger the giant becomes.

Yes, Newton was a genius. That is undeniable. But he was not a Greek god. For all his faults, he displayed characteristics that we should tell our students about, for they are the keys to his, and their, success:

- He built on the best work of the past
- He had brilliant insights
- He worked “by thinking continually”
- He had stubborn perseverance
- He steadily expanded his inquiries

- He made mistakes—and learned from them.

Newton’s success was a synergistic combination of innate genius and immense effort. This is the lesson of history.

References

1. H. J. M. Bos, On the Representation of Curves in Descartes’ *Géométrie*, *Archive History of Exact Sciences* 24 (1981) 295–338.
2. Florian Cajori, *William Oughtred, a Great Seventeenth-Century Teacher of Mathematics*, Open Court, Chicago, 1916.
3. René Descartes, *The Geometry of René Descartes*, translated by D. E. Smith and Marcia L. Latham, Open Court, Chicago, 1925. Reprinted by Dover, 1954.
4. —, *Discourse on Method, Optics, Geometry, and Meteorology*, translated by Paul J. Olscamp, The Library of Liberal Arts, Bobbs-Merrill, Indianapolis, 1965.
5. Betty Jo Teeter Dobbs, *The Foundations of Newton’s Alchemy or “The Hunting of the Green Lyon”*, Cambridge University Press, 1975.
6. Charles Coulston Gillispie (editor), *Dictionary of Scientific Biography*, Scribners, New York, 16 vols, 1970–1980. If there are dates following an individual’s name, then there is a signed article in this work concerning him.
7. Derek Gjertsen, *The Newton Handbook*, Routledge and Kegan Paul, London, 1986.
8. A. W. Grootendorst and J. A. van Maanen, Van Heuraet’s Letter (1659) on the Rectification of Curves. Text, Translation (English, Dutch), Commentary, *Nieuw Archief voor wiskunde* 3:30 (1982) 95–113.
9. A. Rupert Hall, *Philosophers at War: The Quarrel Between Newton and Leibniz*, Cambridge University Press, 1980.
10. John Harrison, *The Library of Isaac Newton*, Cambridge University Press, 1978.
11. Joseph E. Hormann, *Leibniz in Paris, 1672–1676: His Growth to Mathematical Maturity*, Cambridge University Press, 1974.
12. Christiaan Huygens, *The Pendulum Clock or Geometrical Demonstrations Concerning the Motion of Pendula as Applied to Clocks*, translated by Richard J. Blackwell, Iowa State University Press, 1986.
13. Timothy W. Lenoir, *The Social and Intellectual Roots of Discovery in Seventeenth Century Mathematics*, Ph.D. dissertation, Indiana University, 573 pp., 1974. University Microfilms order number 75-1718.

14. Jan A. van Maanen, Hendrick van Heuraet (1634–1660?): His Life and Mathematical Work, *Centaurus* 27 (1984) 218–279.
15. Frank E. Manuel, *A Portrait of Isaac Newton*, Harvard University Press, 1968, and New Republic Books (paperback), Washington D.C., 1979.
16. J. E. McGuire and Martin Tamny, *Certain Philosophical Questions: Newton's Trinity Notebook*, Cambridge University Press, 1983.
17. Isaac Newton, *Sir Isaac Newton's Mathematical Principles of Natural Philosophy and his System of the World*, revision of Andrew Motte's 1729 translation by Florian Cajori, University of California Press, 1934.
18. —, *Isaac Newton's Papers and Letters on Natural Philosophy and Related Documents*, edited by I. B. Cohen, Harvard University Press, 1958.
19. —, *The Correspondence of Isaac Newton*, edited by H. W. Turnbull et al., 7 vols., Cambridge University Press, 1959–1977.
20. —, *The Mathematical Papers of Isaac Newton*, edited by D. T. Whiteside, 8 vols., Cambridge University Press, 1967–1981. References are by volume and page number.
21. —, *The Optical Papers of Isaac Newton*, edited by Alan E. Shapiro, Cambridge University Press, 1984.
22. George Pólya, *Mathematical Discovery*, 2 vols., John Wiley, New York, 1962.
23. J. F. Scott, *The Scientific Work of René Descartes (1596–1650)*, London, 1952.
24. John Wallis, *A Treatise on Algebra Both Historical and Practical . . .*, London, 1685.
25. Robert Weinstock, Dismantling a Centuries-Old Myth: Newton's Principia and Inverse-Square Orbits, *American Journal of Physics*, 50 (1982) 610–617.
26. Richard S. Westfall, Award of the 1977 Sarton Medal to D. T. Whiteside, *Isis* 69 (1978) 86–87.
27. —, Newton's Marvelous Years of Discovery and Their Aftermath: Myth versus Manuscript, *Isis* 71 (1980) 109–121.
28. —, *Never at Rest, A Biography of Isaac Newton*, Cambridge University Press, 1980.
29. —, Newton's Theological Manuscripts, *Contemporary Newtonian Research*, edited by Z. Bechler, D. Reidel, Dordrecht, 1982, pp. 129–143.
30. Derek Thomas Whiteside, Patterns of Mathematical Thought in the later Seventeenth Century, *Archive for History of Exact Sciences* 1 (1961) 179–388.
31. —, Newton's Discovery of the General Binomial Theorem, *Mathematical Gazette* 45 (1961) 175–180.
32. —, Sources and Strengths of Newton's Early Mathematical Thought, *The Annus Mirabilis of Sir Isaac Newton, 1666–1966*, edited by Robert Palter, M.I.T. Press, 1970, pp. 69–85.
33. —, Newton the Mathematician, *Contemporary Newtonian Research*, edited by Z. Bechler, D. Reidel, Dordrecht, 1982, pp. 109–127.
34. —, Newtonian Motion, *Isis* 73 (1982) 100–107. Essay review of [28].

Reading the Master: Newton and the Birth of Celestial Mechanics

BRUCE POURCIAU

American Mathematical Monthly 104 (1997), 1–19

One factor that has remained constant through all the twists and turns of the history of physical science is the decisive importance of the mathematical imagination. — Freeman J. Dyson

1

In January of 1684, the young astronomer Edmund Halley travelled from Islington up to London for a meeting of the Royal Society. Later, perhaps over tea and chocolate at a nearby coffee house, he chatted casually about natural philosophy and other topics with Sir Christopher Wren and Robert Hooke. Talk soon turned to celestial motions, and Halley later reconstructed the conversation [22 p. 26]:

I, having from the consideration of the sesquialter proportion of Kepler concluded that the centripetall force [to the Sun] decreased in the proportion of the squares of the distances reciprocally, came one Wednesday to town, where I met with S^r Christ. Wren and M^r Hook, and falling in discourse about it, M^r Hook affirmed that upon that principle all the Laws of the celestial motions were to be demonstrated, and that he himself had done it. I declared the ill success of my attempts; and S^r Christopher to encourage the Inquiry said that he would give M^r Hook or me 2 months time to bring him a convincing demonstration thereof, and besides the honour, he of us that did it, should have from him a present of a book of 40 shillings. M^r Hook then said that he would conceale [his] for some time that other triing and failing, might know how to value it, when he should make it

publick... I remember S^r Christopher was little satisfied that he could do it, and though M^r Hook then promised to show it him, I do not yet find that in that particular he has been as good as his word.

The two-month deadline passed. Wren and Halley waited through the summer, but still the promised proof from Hooke never came. Finally, in August, Halley would wait on Hooke no longer. He carried the question to Cambridge and the Lucasian Professor of Mathematics, Isaac Newton.

Newton's secretary and attendant has painted a portrait, daubed with colorful and concrete detail, of the eccentric Cambridge professor Halley had finally decided to approach [12, p. xiii-xiv]:

I cannot say, I ever saw him laugh, but once ... I never knew him take an Recreation or Pastime, either in Riding out to take y^e Air, Walking, Bowling or any other Exercise whatever, thinking all Hours lost, y^t was not spent in his Studies, to w^{ch} he kept so close ... so intent, so serious upon [them], y^t he eat very sparingly, nay, oft times he has forgot to eat at all, so y^t going into his Chamber I have found his Mess untouch'd, of w^{ch} when I have reminded him, [he] would reply, Have I; & then making to y^e Table, would eat a bit or two standing, for I cannot say, I ever saw Him sit at Table by himself ... He very rarely went to Dine in y^e Hall unless upon some Publick Dayes, & then, if He has not been minded, would go very carelessly, wth Shoes down at Heels, Stockins unty'd, Suplice on, & his Head scarcely comb'd ... At some seldom Times when he design'd to dine

in y^e Hall [he] would turn to y^e left hand, & go out into y^e street, where making a Stop, when he found his mistake, [he] would hastily turn back & then sometimes instead of going into y^e Hall, would return to his Chamber again . . .

. . . in his Garden, w^{ch} was never out of Order, . . . he would, at some seldom Times, take a short Walk or two, not enduring to see a Weed in it . . . When he has some Times taken a turn or two [he] has made a sudden Stand, turn'd himself about, run up y^e Stairs [&] like another A[r]chimides, with an *εὐρηκα* fall to write on his Desk standing, without giving himself the Leisure to draw a Chair to sit down on . . .

In a letter from 1727 [22, p. 27], Abraham de Moivre set the scene as Halley, having arrived in Cambridge, posed the crucial question to the reclusive mathematician:

. . . after they had been some time together, the D^r asked [Newton] what he thought the Curve would be that would be described by the Planets supposing the force of attraction towards the Sun to be reciprocal to the square of their distance from it. S^r Isaac replied immediately that it would be an Ellipsis. The Doctor struck with joy and amazement asked him how he knew it. Why saith he I have calculated it . . .

Witness the birth of celestial mechanics: the embryonic question has been answered—

every orbital motion subject to an inverse-square force lies on a conic having focus at the force center

—not with a guess, but with a *mathematical demonstration!*

Semester after semester, at every college and university, we give our students the same answer Newton gave to Halley, our demonstrations—so different from Newton's—blessed by the glories of vector calculus, and in this way we honor Newton and celebrate the emergence of celestial dynamics. In the present article, we honor Newton in the way of Abel, who counsels us to read the masters. We shall place the original argument from Newton's *Principia* next to a modern counterpart, delighting in the stark contrasts. One delightful difference: Newton's argument requires that we first answer the converse to Halley's question—

What force law maintains a conic motion orbiting about the focus?

—and again, reading the master, we shall juxtapose the *Principia's* very geometric proof of this reversal with its demonstration by vector calculus. In this mix of old and new, of geometry and analysis, some insights and surprises make their way to the surface:

- The mathematics of the *Principia* is geometric *analysis*, both analysis in the sense of 'taking apart' as well as analysis in the sense of *calculus*. Newton's geometry is calculus—limits, derivatives, integrals, acceleration, curvature—masked as geometry.
- While less precise than their vector calculus descendants, the *Principia's* definitions have a concrete, visceral character that informs our geometric and physical intuition.
- The first ten sections of the *Principia* (apart from the statement of the Third Law) contain no physics, only mathematics. Newton may write of 'forces,' but he calculates accelerations. His concentration on acceleration and shape reminds us that force and mass take no part in the mathematics of the one-body problem, which occupies the leading sections of the *Principia*.
- In contrast to force, curvature is deeply involved with the *Principia's* orbital dynamics, yet apart from rare oblique sightings, the dependence on curvature remains hidden.
- Asked who should receive credit for answering Halley's question with a demonstration rather than a guess, historians of science bow to Newton. Asked for evidence to back up their claim, the historians open the *Principia* and point to a *two-sentence* argument. We confirm that Newton's little sketch, given air and sun, blossoms into a cogent proof.
- Reading the masters—Archimedes, Newton, Euler, Gauss, Riemann, . . .—can mean entering a foreign paradigm, an unfamiliar mathematical world where alien values, language, definitions, tools, strategies, and assumptions frustrate our attempts to understand. And so it is with the *Principia*. But with persistence and prayer, even the *Principia* sends up her secrets. As we slowly learn to navigate in Newton's world, we deepen our understanding of the *Principia's* paradigm as well as our own.

It may seem odd to have placed our conclusions here in the introduction, but with these closing remarks

now out of the way, we can read on unburdened by the western need to fret and fuss about the point of it all. As the Taoist philosopher Chuang Tzu suggests [19, p. 126], we can now lean back and float with the current, “going under with the swirls and coming out with the eddies, following along the way the water goes, and never thinking . . .”

2

We begin with Newton’s generalized answer to Halley—that every orbit produced by an inverse-square force must lie on a conic—in this section giving a contemporary proof and in the next exploring the *Principia*’s original argument. But we should first agree on some technical vocabulary, so that we can be more precise. Any smooth map $\mathbf{r} = \mathbf{r}(t)$ from an open interval J into euclidean 3-space is a *motion*. Every motion \mathbf{r} has a velocity $\mathbf{v} = \dot{\mathbf{r}}$ and an acceleration $\mathbf{a} = \dot{\mathbf{v}}$. For the magnitude of a vector, we choose the same letter in non-bold italic: thus, for example, $r = |\mathbf{r}|$, $v = |\mathbf{v}|$, and $a = |\mathbf{a}|$. (We tacitly assume that r and v (the *speed*) never vanish.) We say the motion \mathbf{r} has an *inverse-square acceleration* provided for some non-zero λ ,

$$\mathbf{a} = \frac{-\lambda}{r^2} \mathbf{U}$$

for all t in J . Here \mathbf{U} stands for the unit direction vector \mathbf{r}/r . More generally, whenever the cross-product $\mathbf{r} \times \mathbf{a}$ vanishes identically, we call \mathbf{r} an *orbital motion*. If the origin S has some significance—it might be the focus of a conic or the pole of a spiral, for instance—an orbital motion may be labelled a *motion about S*. A sentence that would be typical of the *Principia*, “A body is urged by a centripetal force continually directed toward an immovable center S ,” becomes briefer in our language: “Given a motion about S .”

Assuming that Mars traversed an ellipse with its position vector sweeping out equal areas in equal times, Kepler made predictions in his *Astronomica nova* of 1609 that matched the careful observations of Tycho Brahe. In Propositions I and II (Section II, Book 1) of the *Principia*, Newton uses this area principle to characterize orbital motions in general [11, p. 40 and 42]:

Proposition I Theorem I

The areas which revolving bodies describe by radii drawn to an immovable center of force do lie in the same immovable planes, and are

proportional to the times in which they are described.

Proposition II Theorem II

Every body that moves in any curved line described in a plane, and by a radius drawn to a point either immovable, or moving forwards with an uniform rectilinear motion, describes about that point areas proportional to the times, urged by a centripetal force directed to that point.

Today of course we translate these propositions into the language of vectors:

Newton’s Area Theorem. *For any motion $\mathbf{r} = \mathbf{r}(t)$, the following are equivalent:*

- (a) \mathbf{r} is orbital
- (b) the (massless) angular momentum $\mathbf{h} = \mathbf{r} \times \mathbf{v}$ is constant.
- (c) \mathbf{r} is planar and sweeps out area at a constant rate.

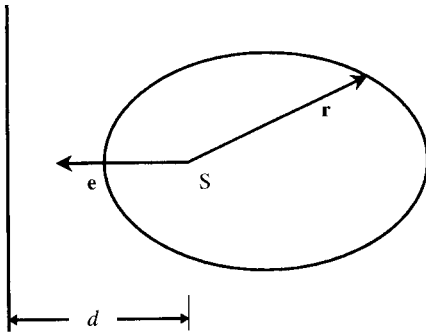
The proof is simple, especially once we agree that the area swept out is

$$\frac{1}{2} \int_{t_0}^t |\mathbf{r} \times \mathbf{v}| dt,$$

the only slippery step being to show \mathbf{r} is planar when \mathbf{h} vanishes everywhere, but in this case the derivative $\dot{\mathbf{U}}$ vanishes everywhere (recall $\mathbf{U} = \mathbf{r}/r$), indicating that the motion lies on a fixed ray from the origin. That $\dot{\mathbf{U}}$ remains zero follows from a simple fact:

$$\dot{\mathbf{U}} = \frac{\mathbf{h} \times \mathbf{r}}{r^3} \tag{1}$$

Halley’s question and Newton’s answer involve the relationship between the acceleration of the motion and the shape of the orbit. Moving from acceleration to shape, we define the *trajectory* of a motion $\mathbf{r} = \mathbf{r}(t)$ to mean the subset $\{\mathbf{r}(t) : t \in J\}$ of 3-space. An *orbit* is then just the trajectory of an orbital motion. If a trajectory lies on a conic, say, or a spiral, we would have a *conic* or *spiral motion*. The Principian sentence, “A body, urged by a centripetal force continually directed toward an immovable center S , moves in a conic section with focus at S ,” now turns into “Consider a conic motion about S .” Of course conics hold some special interest for us here, and we recall the following definition: a *conic* is the locus of points whose distance from a given point S (the *focus*) is some positive



constant e (the *eccentricity*) times the distance from a given line (the *directrix*). Perhaps we should put this definition in vector dress, so it will feel more comfortable when vector calculus comes to call. If we let \mathbf{r} be the position vector from the focus, d the distance from the directrix to the focus, and \mathbf{e} (the *eccentricity vector*) a vector of length e which points perpendicularly toward the directrix, then the definition tells us that

$$r = e \left(d - \mathbf{r} \cdot \frac{\mathbf{e}}{e} \right),$$

and with the notation $\mathbf{U} = \mathbf{r}/r$ and $l = de$, this formula turns into the *vector conic equation*:

$$\mathbf{r} \cdot (\mathbf{e} + \mathbf{U}) = l. \quad (2)$$

The constant l is called the *semi-latus rectum* of the conic. Given a positive constant l and a non-zero vector \mathbf{e} , the vector conic equation defines a conic with semi-latus rectum l , eccentricity $e = |\mathbf{e}|$, axis along \mathbf{e} , and focus at the origin. When $\mathbf{e} = \mathbf{0}$, then (2) describes a circle of radius l about the origin, and if $l = 0$, we have a ray from the origin.

At this point, we have the vocabulary and background to explore a contemporary version of Newton's answer to Halley. Suppose we have a motion $\mathbf{r} = \mathbf{r}(t)$ with an inverse-square acceleration, so that for some non-zero number λ ,

$$\mathbf{a}(t) = \frac{-\lambda}{r^2} \mathbf{U}(t)$$

for all t in some open interval J . Crossing with the angular momentum $\mathbf{h} = \mathbf{r} \times \mathbf{v}$, we have

$$\mathbf{a} \times \mathbf{h} = \frac{-\lambda}{r^2} \mathbf{U} \times \mathbf{h} = -\lambda \frac{\mathbf{r} \times \mathbf{h}}{r^3}$$

which becomes, using (1),

$$\mathbf{a} \times \mathbf{h} = \lambda \dot{\mathbf{U}}.$$

Now antidifferentiate, remembering that \mathbf{h} is constant because \mathbf{r} is orbital:

$$\mathbf{v} \times \mathbf{h} = \lambda \mathbf{U} + \mathbf{c} = \lambda(\mathbf{U} + \mathbf{e})$$

for some constant vectors \mathbf{c} and $\mathbf{e} = (1/\lambda)\mathbf{c}$. If we dot with \mathbf{r} , we find

$$\frac{1}{\lambda} \mathbf{r} \cdot (\mathbf{v} \times \mathbf{h}) = \mathbf{r} \cdot (\mathbf{e} + \mathbf{U}),$$

and then permuting the entries in the scalar triple product uncovers the vector conic equation (2):

$$\frac{h^2}{\lambda} = \mathbf{r} \cdot (\mathbf{e} + \mathbf{U}).$$

When the constant vector \mathbf{h} vanishes, this reduces to $\mathbf{U} = -\mathbf{e}$, and the motion must then lie on a fixed ray from the origin. If \mathbf{h} does not vanish, but \mathbf{e} does, we conclude $r = h^2/\lambda$, so the orbit lies on a circle centered at the origin. Supposing neither \mathbf{h} nor \mathbf{e} vanishes, we have seen that the vector conic equation (2) defines a conic with focus at the origin. And that seals it:

Newton's Shape Theorem. *Apart from motion on a ray from the center, every motion with an inverse-square acceleration must be a conic motion about the focus.*

A second proof of the Shape Theorem is quick but sly. Assume again that

$$\mathbf{a}(t) = \frac{-\lambda}{r^2} \mathbf{U}(t).$$

Then of course \mathbf{h} remains constant, but (surprise!) so does the vector

$$\mathbf{L} = \frac{1}{\lambda} \mathbf{v} \times \mathbf{h} - \mathbf{U}.$$

To check, compute the derivative:

$$\begin{aligned} \dot{\mathbf{L}} &= \frac{1}{\lambda} \mathbf{a} \times \mathbf{h} - \frac{\mathbf{h} \times \mathbf{r}}{r^3} \\ &= \frac{1}{\lambda} \left(\frac{-\lambda}{r^2} \mathbf{U} \right) \times \mathbf{h} - \frac{\mathbf{h} \times \mathbf{U}}{r^2} = \mathbf{0}. \end{aligned}$$

Now just dot \mathbf{r} with $\mathbf{L} + \mathbf{U}$,

$$\mathbf{r} \cdot (\mathbf{L} + \mathbf{U}) = \frac{1}{\lambda} \mathbf{r} \cdot (\mathbf{v} \times \mathbf{h}) = \frac{h^2}{\lambda},$$

and we recognize the vector conic equation (2). That's all there is to it.

The sly part of this proof is (un)clear: why would one *expect* the vector $\frac{1}{\lambda} \mathbf{v} \times \mathbf{h} - \mathbf{U}$ to be constant? The

secret lies in a formula for the eccentricity vector \mathbf{e} . Given any conic motion $\mathbf{r} = \mathbf{r}(t)$, if we differentiate the vector conic equation,

$$\mathbf{r} \cdot (\mathbf{e} + \mathbf{U}) = l,$$

and solve for the (constant) eccentricity vector \mathbf{e} , we obtain the

Eccentricity Formula. For any motion $\mathbf{r} = \mathbf{r}(t)$ satisfying the vector conic equation (2),

$$\mathbf{e} = \frac{l}{h^2} \mathbf{v} \times \mathbf{h} - \mathbf{U}. \quad (3)$$

Of course we began with an inverse-square motion, not a conic motion, but if we had had a conic motion, then the vector $(l/h^2)\mathbf{v} \times \mathbf{h} - \mathbf{U}$, representing as it does the eccentricity vector, would have been *a priori* constant. Knowing that λ turns out to be h^2/l (see our first proof), it seems natural then to suspect that $\mathbf{L} = (1/\lambda)\mathbf{v} \times \mathbf{h} - \mathbf{U}$ should be constant in the case of inverse-square acceleration. If you do not like this sneaky proof of the Shape Theorem, blame Laplace. The vector \mathbf{L} , sometimes called the Laplace-Runge-Lenz vector, has the history of its rediscoveries etched in its name.

Now that we have seen two contemporary proofs, let us drift back in time, back to the 1680s, to examine Newton's original argument for the Shape Theorem in the *Principia*.

3

Only with some nervousness, do we open Newton's monumental work *Philosophiæ Naturalis Principia Mathematica*. It had a reputation in 1687; it has a reputation still—a reputation for being impenetrable. In the latter half of the eighteenth century and on into the nineteenth, this reputation fed a cottage industry of writing notes and commentaries devoted entirely to ‘understanding’ the *Principia*. (The industry may have declined, but it still produces excellent commentaries from time to time: witness [5] and [6], just out in 1995.) Always formal, terse, and crabbed in his scholarly work, Newton took these stylistic tendencies to their limit in the *Principia*. Why? A decade earlier, his theory of colors had been attacked by Leibniz, Hooke, Linus, Lucas, as well as others, and Newton had detested the controversy. In a shrill letter to Henry Oldenburg, who was then Secretary of the Royal Society, Newton despairs, “I see I have made myself a slave to Philosophy, but if I get free of Mr. Linus's business I will resolutely bid adew

to it eternally, excepting what I do for my private satisfaction or leave to come out after me. For I see a man must either resolve to put out nothing new or become slave to defend it.” [7, p. 198] Of course, Newton did not “leave [the *Principia*] to come out after [him],” but he did choose to limit his readership and therefore his potential critics by composing in an icy, mathematical style, ultimately producing 500 pages of dense Latin text—definitions, axioms, lemmas, theorems, propositions, demonstrations, scholia, and figures, all fixed in place, a massive ordered regiment of abstract formality. According to a close friend of Newton's [2, p. 168] controversy of any kind

made sr Is[aac] very uneasy; who abhorred all Contests . . . And for this reason, mainly to avoid being baited by little Smatterers in Mathematicks, he told me, he designedly made his *Principia* abstruse; but yet so as to be understood by able Mathematicians, who he imagined, by comprehending his Demonstrations would concurr with him in his Theory.

Yet even the most able mathematicians of the day struggled with the *Principia*. The confident young mathematician Abraham de Moivre happened to be visiting the Duke of Devonshire when Newton arrived to present the Duke with a copy of the new work [21, p. 471]:

[de Moivre] opened the book and deceived by its apparent simplicity persuaded himself that he was going to understand it without difficulty. But he was surprised to find it beyond the range of his knowledge and to see himself obliged to admit that what he had taken for mathematics was merely the beginning of a long and difficult course that he had yet to undertake. He purchased the book, however; and since the lessons he had to give forced him to travel about continually, he tore out the pages in order to carry them in his pocket and to study them during his free time.

Prepared by its scary reputation, we cannot conjure up the initial poise of de Moivre as we open the *Principia*, but prepared for some hard work, let us take a look at Newton's argument for the Shape Theorem. Actually, to do this in the proper order, we should close the *Principia* for the moment and begin nearer the beginning, returning to Halley's call on Newton in 1684. Earlier we have read de Moivre's description of their meeting [22, p. 27]:

... after they had been some time together, the D^r asked him what he thought the Curve would be that would be described by the Planets supposing the force of attraction towards the Sun to be reciprocal to the square of their distance from it. S^r Isaac replied immediately that it would be an Ellipsis. The Doctor struck with joy and amazement asked him how he knew it. Why saith he I have calculated it ...

But stopping here is a rude interruption, for de Moivre continues [7, p. 283],

... whereupon D^r Halley asked him for his calculation without any farther delay, S^r Isaac looked among his papers but could not find it, but he promised him to renew it, & sent it.

It would be three months before Newton made good his promise, but idleness had not caused the delay, for he not only renewed his calculation for the ellipse, but embedded that calculation in a nine-page tract, *De motu Corporum in gyrum (On the Motion of Bodies in Orbit)*, which Halley received in November.

It is in *De motu* then that we should look for Newton's original demonstration of the Shape Theorem, that an inverse-square force implies conic orbits. Thumbing through its pages, we pass a line of definitions, hypotheses, theorems, corollaries, and problems until we stop at a familiar-looking claim [12, VI, 49]:

Scholium *The major planets orbit, therefore, in ellipses having a focus at the centre the Sun ... exactly as Kepler supposed.*

The Shape Theorem (at least for ellipses)! Eagerly we anticipate the proof — hunched over the scholium, eyes narrowed, pencil poised — but then the adrenaline seeps away as we scan down the page to find ... nothing. Newton has left the Shape Theorem, his answer to Halley, as a bald claim, completely unsupported! Because the scholium directly follows:

Problem 3 *A body orbits in an ellipse: there is required the law of centripetal force tending to a focus of the ellipse.*

we would guess that Newton must have viewed the Shape Theorem as a trivial corollary of his solution to Problem 3, or, more generally, of what we shall call

Newton's Acceleration Theorem. *Every conic motion about the focus has an inverse-square acceleration.*

Not understanding how the Shape Theorem would follow trivially from the Acceleration Theorem, we turn from *De motu* to the *Principia*, expecting the fuller development there to enlighten us.

Halley's question in August of 1684 had reseeded Newton's interest in celestial mechanics, and *De motu* was just the first little sprout. In January of 1685, he wrote Flamsteed, the Astronomer Royal, "Now that I am upon this subject, I would gladly know ye bottom of it before I publish my papers." [7, p. 286]. What understatement: between November of 1684 and April of 1687, Newton came to "know ye bottom of it," and the nine-page treatise exploded into a five hundred page masterpiece.

Now remember that *De motu* had left the Shape Theorem unproved. And the 1687 *Principia*? No better! In Section III of Book I, Newton demonstrates Propositions XI-XIII, which, taken together, establish the Acceleration Theorem and then follows with the Shape Theorem dressed as a corollary [11, p. 61] to this trio of propositions:

Cor. I. *From the three last Propositions it follows, that if a body P goes from place P with any velocity in the direction of any right line PR, and at the same time is urged by the action of a centripetal force that is inversely proportional to the square of the distance of the places from the center, the body will move in one of the conic sections, having its focus in the center of force ...*

But again, no proof. Worse yet, no one complained — not Halley, not Leibniz, not Huygens, not de Moivre — until, in October of 1710, twenty-three years after the publication of the *Principia*, Johann Bernoulli finally pointed out the obvious: Corollary I needed a demonstration. By this time, however, perhaps getting an early wind of Bernoulli's criticism, Newton had already decided to fill the gap, instructing his editor, in a letter dated 11 October 1709, to slip the following argument [13, p. 5–6] into the second edition (1713) of the *Principia*:

Nam datis umbilico et puncto contactus & positionis tangentis, describi potest Sectio conica quae curvaturam datam ad punctum illud habebit. Datur autem curvature ex data vi centripeta: et Orbes duo se mutuo tangentes eadem vi describi non possunt.

For the third edition (1726), Newton added to this shockingly brief sketch the word ‘velocity’ in two places, resulting in [11, p. 61]

Newton’s Argument for the Shape Theorem

For the focus, the point of contact, and the position of the tangent, being given, a conic section may be described, which at that point shall have a given curvature. But the curvature is given from the centripetal force and velocity of the body be given; and two orbits, touching one the other, cannot be described by the same centripetal force and the same velocity.

Brevity may be the soul of wit, but it may be the seed of confusion as well. No one laughs when a fundamental proposition of celestial mechanics is followed by a two-sentence sketch which fails to persuade. At least Newton’s *plan*, although strikingly different from what we saw in Section 2, seems both familiar and clear—to prove that every solution to a given initial-value problem has a particular form, we exhibit a solution of that form and then invoke a uniqueness principle—but connecting all the dots in the outline may be another story especially when some of the dots themselves are missing.

Expanding Newton’s sketch in a natural way, we arrive at what we take as his intended strategy:

Newton’s Strategy for Proving the Shape Theorem

1. Suppose given any motion $\bar{\mathbf{r}} = \bar{\mathbf{r}}(t)$ with an inverse-square acceleration. At some time t_0 , note the position \mathbf{r}_0 , velocity \mathbf{v}_0 , and curvature κ_0 of the motion $\bar{\mathbf{r}}$.
2. Construct a conic \mathcal{C} , having focus at the origin, that passes through the tip of \mathbf{r}_0 with tangent parallel to \mathbf{v}_0 and curvature κ_0 .
3. On the conic \mathcal{C} , put a motion $\mathbf{r} = \mathbf{r}(t)$ about the focus that leaves the tip of \mathbf{r}_0 with velocity \mathbf{v}_0 . (Newton never mentions this step, which involves making sure the position vector sweeps out area at a uniform rate, but it’s a simple matter, and one that he probably took for granted.)
4. From Propositions XI–XIII (the Acceleration Theorem), infer that $\mathbf{r} = \mathbf{r}(t)$, a conic motion about the focus, must have an inverse-square acceleration.
5. Thus both \mathbf{r} and $\bar{\mathbf{r}}$ have inverse-square accelerations, but even better, the matching of position,

velocity, and curvature in steps (2) and (3) forces \mathbf{r} and $\bar{\mathbf{r}}$ to share the same proportionality constant.

6. Finally, noting that \mathbf{r} and $\bar{\mathbf{r}}$ now both solve the same initial-value problem, invoke a uniqueness principle to conclude that $\mathbf{r} = \bar{\mathbf{r}}$, proving that our given inverse-square motion \mathbf{r} must be a conic motion about the focus as desired.

As we begin to check whether this six-step strategy unfolds further into a convincing proof, we can see already that step (2) will block us, unless we know a little about the curvature of conics. For a motion $\mathbf{r} = \mathbf{r}(t)$, the curvature κ is $|\dot{\mathbf{T}}|/v$ and the radius of curvature ρ is $1/\kappa$, where \mathbf{T} is the unit tangent \mathbf{v}/v . From the velocity and the acceleration, we can easily find the curvature from a well-known formula:

$$\rho = \frac{v^3}{|\mathbf{a} \times \mathbf{v}|}. \tag{4}$$

To calculate the radius of curvature for a conic, we start with any motion $\mathbf{r} = \mathbf{r}(t)$ satisfying the vector conic equation (2),

$$\mathbf{r} \cdot (\mathbf{e} + \mathbf{U}) = l,$$

differentiate twice to get

$$\mathbf{a} \cdot (\mathbf{e} + \mathbf{U}) + \mathbf{v} \cdot \frac{\mathbf{h} \times \mathbf{r}}{r^3} = 0,$$

and insert our formula (3) for the eccentricity vector \mathbf{e} to see that

$$\frac{l}{h^2} \mathbf{a} \cdot (\mathbf{v} \times \mathbf{h}) + \mathbf{v} \cdot \frac{\mathbf{h} \times \mathbf{r}}{r^3} = 0.$$

Sliding the entries in the scalar triple products gives back

$$|\mathbf{a} \times \mathbf{v}| = \frac{1}{l} \left(\frac{h}{r} \right)^3,$$

which leads to

$$\rho = \frac{v^3}{|\mathbf{a} \times \mathbf{v}|} = l \left(\frac{rv}{h} \right)^3,$$

or, rephrasing, to the

Conic Curvature Lemma. *For any conic motion with semi-latus rectum l ,*

$$\rho = \frac{l}{|\mathbf{U} \times \mathbf{T}|^3}. \tag{5}$$

Newton cast this lemma more elegantly [12, III p. 159]: *If the line perpendicular to the conic at P meets the focal axis at N , then ρ varies as PN^3 .*

(The equivalence to our lemma follows from a geometric fact about conics: the projection of PN onto SP is the semi-latus rectum.) This lovely property is just one of several striking results on curvature obtained by Newton in his 1671 tract on series and fluxions. “The problem [of curvature],” he wrote in this tract, “has the mark of exceptional elegance and of being pre-eminently useful in the science of curves.” [12, III p. 151] From an insight in his *Waste Book* made around December of 1664 (over twenty years before the *Principia*), we have evidence that Newton also recognized the fundamental place of curvature in the study of orbital motions: “If the body b moved in an Ellipsis, then its force in each point (if its motion in that point bee given) may be found by a tangent circle of equall crookedness [read curvature] with that point of the Ellipsis.” [22, p. 14] It is perhaps surprising then that curvature plays no role in the 1687 *Principia*. However, in the 1690s Newton made radical plans for revising the first edition, plans that would have made curvature the centerpiece of his celestial mechanics. Sadly, this radical revision never made it into print, and in the end Newton contented himself with relatively minor changes, squeezing some curvature methods into the second (1713) and third (1726) editions as tacked-on corollaries. For more on the role of curvature in Newton’s celestial mechanics, see [3, 4, 10, and 17].

Now that we know something about the curvature of conics, we can begin to connect all the dots in a proof of the Shape Theorem inspired by Newton’s two-sentence argument in the *Principia*. We follow the six-step strategy above, for it seems to be the one plausible interpretation of what Newton had in mind.

Step 1: We give ourselves any motion $\bar{\mathbf{r}} = \bar{\mathbf{r}}(t)$ with an inverse-square acceleration: for some nonzero λ , suppose $\bar{\mathbf{r}}$ solves the initial-value problem

$$\ddot{\bar{\mathbf{r}}}(t) = \frac{\lambda}{r^2} \mathbf{U}(t), \quad \bar{\mathbf{r}}(t_0) = \mathbf{r}_0, \quad \dot{\bar{\mathbf{r}}}(t_0) = \mathbf{v}_0$$

on the open interval J . If $\mathbf{r}_0 \times \mathbf{v}_0 = \mathbf{0}$, then the motion lies on a fixed ray through the origin, but apart from this special case, we need to prove that $\bar{\mathbf{r}}$ is a conic motion about the focus. Since $\bar{\mathbf{r}}$ is an orbital motion, the orbit lies in a fixed plane and the angular momentum remains fixed at $\mathbf{h}_0 = \mathbf{r}_0 \times \mathbf{v}_0$.

Step 2: In this fixed plane, we now construct a conic that “fits” the orbit of $\bar{\mathbf{r}}$. Let ρ_0 the radius of curvature of $\bar{\mathbf{r}}$ at $\bar{\mathbf{r}}(t_0) = \mathbf{r}_0$. Put

$$l = \rho_0 |\mathbf{U}_0 \times \mathbf{t}_0|^3, \quad \mathbf{e} = \frac{l}{h_0^2} \mathbf{v}_0 \times \mathbf{h}_0 - \mathbf{U}_0$$

where $\mathbf{U}_0 = \mathbf{r}_0/r_0$, $\mathbf{T}_0 = \mathbf{v}_0/v_0$, and $\mathbf{h}_0 = \mathbf{r}_0 \times \mathbf{v}_0$. (As \mathbf{r}_0 and \mathbf{v}_0 are not parallel, $\mathbf{h}_0 \neq \mathbf{0}$ and \mathbf{e} is well-defined.) The vector-conic equation (2)

$$\mathbf{r} \cdot (\mathbf{e} + \mathbf{U}) = l$$

now defines a particular conic \mathcal{C} . One easily checks that \mathcal{C} has a focus at the origin, and that \mathcal{C} passes through the tip of \mathbf{r}_0 with its tangent parallel to \mathbf{v}_0 and its radius of curvature equal to ρ_0 .

Step 3: At this point, we would like to apply Newton’s Acceleration Theorem to our constructed conic, but the Acceleration Theorem applies only to conic motions, indeed only to conic motions about the focus, not to mere conic loci. Therefore, on the conic locus \mathcal{C} we now place a motion about the focus. (To put it differently, we must parameterize the conic locus \mathcal{C} in a way that keeps the acceleration vector pointed at the focus.) By the Area Theorem, to make a motion about the focus, we need only make a motion whose position vector from the focus sweeps out area at a constant rate, and intuitively we can do this by arranging for the area swept out to be our parameter. More precisely: Using arc-length measured from the tip of \mathbf{r}_0 , let $\mathbf{r}_1 = \mathbf{r}_1(s)$ be the unit-speed motion on \mathcal{C} having initial velocity \mathbf{T}_0 . The real function

$$a(s) = t_0 + \int_0^s \frac{1}{h_0} |\mathbf{r}_1(s) \times \dot{\mathbf{r}}_1(s)| ds$$

is smooth and strictly increasing. (Note that $h_0 = |\mathbf{r}_0 \times \mathbf{v}_0| \neq 0$ and $|\mathbf{r}_1(s) \times \dot{\mathbf{r}}_1(s)| \neq 0$ for all s , because tangents to \mathcal{C} never pass through the focus.) Take the (smooth) inverse $a^{-1} = a^{-1}(t)$, and use it to define a motion $\mathbf{r} = \mathbf{r}(t)$ on \mathcal{C} by

$$\mathbf{r}(t) = \mathbf{r}_1 [a^{-1}(t)].$$

This constructed conic motion \mathbf{r} is also a motion about the focus S , for it has constant angular momentum $\mathbf{h}_0 = \mathbf{r}_0 \times \mathbf{v}_0$. Moreover, $\mathbf{r}(t_0) = \mathbf{r}_0$ and $\dot{\mathbf{r}}(t_0) = \mathbf{v}_0$.

We haven’t done anything here, by the way, that Newton couldn’t do. You can find him geometrically constructing motions about the focus, on given conic loci, in the *Principia*, Book I, Section VI [11, p. 109–116]. Such constructions are even implicit in Newton’s proof of the Area Theorem in Propositions I and II, at the very beginning of the *Principia*. In his two-sentence argument for the Shape Theorem, Newton fails to mention the problem of putting an orbital motion on his constructed conic, but at the *Principia*’s level of rigor, this is a trivial omission.

Refer to [15 and 16] for some discussion of this point.

Step 4: We apply the Acceleration Theorem (Propositions XI–XIII, Section III, Book 1) to $\mathbf{r} = \mathbf{r}(t)$, our newly minted conic motion about the focus, and conclude that \mathbf{r} has an inverse-square acceleration: for some nonzero μ ,

$$\ddot{\mathbf{r}}(t) = \frac{\mu}{r^2} \mathbf{U}(t)$$

for all t .

Step 5: To prove that $\mu = \lambda$, we return to the curvature matching we did in Step 2. By design, both our constructed motion \mathbf{r} and our given motion $\bar{\mathbf{r}}$ share the same radius of curvature at the tip of \mathbf{r}_0 , namely ρ_0 . For the conic motion \mathbf{r} , by (4),

$$\rho_0 = \frac{v_0^3}{|\mathbf{a}_0 \times \mathbf{v}_0|} = \frac{v_0^3}{\left| \frac{\mu}{r_0^2} \mathbf{U}_0 \times \mathbf{v}_0 \right|} = \frac{h_0^2}{\mu |\mathbf{u}_0 \times \mathbf{T}_0|^3}.$$

Similarly, for the given motion $\bar{\mathbf{r}}$,

$$\rho_0 = \frac{h_0^2}{\lambda |\mathbf{U}_0 \times \mathbf{T}_0|^3}.$$

It follows that $\mu = \lambda$.

Step 6: We now have two solutions, the constructed conic motion \mathbf{r} and the given inverse-square motion $\bar{\mathbf{r}}$ to the initial-value problem

$$\ddot{\mathbf{r}}(t) = \frac{\lambda}{r^2} \mathbf{U}(t), \quad \mathbf{r}(t_0) = \mathbf{r}_0, \quad \dot{\mathbf{r}}(t_0) = \mathbf{v}_0$$

on the interval J . By standard uniqueness theorems (equivalent to Propositions XLI and XLII, Section VIII, Book I, *Principia*) for differential equations, we conclude that $\mathbf{r} = \bar{\mathbf{r}}$ on J , and it follows that our given inverse-square motion must be a conic motion about the focus, as expected.

This completes a “Newtonian” proof of the Shape Theorem—that every motion having an inverse-square acceleration is a conic motion about the focus—a proof springing from Newton’s two-sentence argument in the *Principia*. Is this proof the contemporary version of what Newton had in mind? Probably, but the sheer brevity of his sketch leaves room for other views. On this issue, read [15, 16, 20, and 23].

Of course, our “completed” Newtonian demonstration is really anything but complete, since in step four, to ensure that our constructed conic motion had an inverse-square acceleration, we called on the *unproved* reversal of the Shape Theorem:

Newton’s Acceleration Theorem. *Every conic motion about the focus has an inverse-square acceleration.*

We now intend to study the original argument for the Acceleration Theorem and then contrast the original with what we might do today, but as we return with this intention to the *Principia* (and specifically to Propositions XI, XII, and XIII in Book I), we must first page back to Proposition VI in order to understand how Newton measures orbital acceleration.

4

In May of 1686, just one month after the *Principia* was presented to the Royal Society, Halley sent news to Newton of the plans for printing and publication, but his cheerful letter ended with a sour lemon [21, p. 446]: “There is one thing more I ought to informe you of,” he wrote,

that M^r Hook has some pretensions upon the invention of y^e rule of the decrease of Gravity, being reciprocally as the squares of the distances from the Center. He sais you had the notion from him . . . how much of this is so, you know best, as likewise what you have to do in this matter, only M^r Hook seems to expect you should make some mention of him, in the preface . . .

“Now is not this very fine?” sneered back Newton [21, p. 448],

Mathematicians that find out, settle & do all the business must content themselves with being nothing but dry calculators & drudges & another that does nothing but pretend & grasp at all things must carry away all the invention . . . And why should I record a man for an Invention who founds his claim upon an error therein & on that score gives me trouble? He imagine he obliged me by telling me his Theory, but I thought myself disobliged by being upon his own mistake corrected magisterially & taught a Theory w^{ch} every body knew & I had a truer notion of then himself.

In his fury at Hooke’s pretensions, Newton struck back with his pen, literally striking out almost every reference to Hooke in the entire *Principia*.

Even so, Hooke did in fact make one significant contribution to the *Principia* for he was the first to see orbital motions as the geometric signature of a

central attraction that pulls the orbiting body away from its linear inertial path. In November of 1679 as the new Secretary of the Royal Society, Hooke had asked Newton to [22, p. 22] “please . . . continue your former favors to the Society communicating what shall occur to you that is Philosophicall,” and he added,

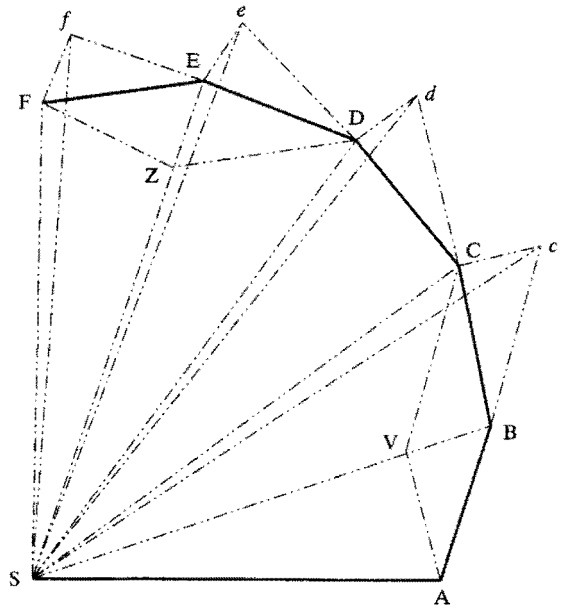
for my own part I shall take it as a great favor . . . if you will let me know your thoughts of [my hypothesis] of compounding the celestial motions of the planets of a direct [straight] motion by the tangent & an attractive motion towards the centrall body.

Hooke had this hypothesis as early as 1670, a time when Newton’s eyes were still clouded by thoughts of “outward endeavor” and “Cartesian vortices.” Still, Hooke’s physical insight could take him only so far. In his hands, the hypothesis remained just that: a guess, a guess rooted in physical intuition and mechanical experiment, yet still a guess. But in Newton’s hands, the hands of a soaring mathematical imagination, Hooke’s hypothesis rose to an aerie of definitions, lemmas, and propositions. Look, for example, at the figure Newton draws to illustrate his proof of Propositions I and II (Section II, Book I), where we see, for the very first time, the mathematical equivalence of central attraction and the area law, and you behold, in its central attraction and deviations from the tangent, the risen form of Hooke’s hypothesis.

Later, in Proposition VI, Newton fashions from Hooke’s inward deviation a formula for measuring the acceleration of an orbital motion. (In the *Principia*, accelerations for *general* motions are never even defined.) If a particle in orbital motion falls freely toward the acceleration center *S*, Newton may have reasoned that the particle could be thought of as instantaneously in free fall from the tangent down to its position on the orbit. In a given time *t*, suppose a particle moves along its orbit from *P* to *Q*. If there had been no acceleration during this time interval, the particle would have proceeded instead along the tangent at constant speed *v* to a location *L*. The deviation *QL*, nearly parallel to *SP*, would be like the “distance fallen toward *S*,” which we would expect to be approximately $\frac{1}{2}at^2$, where *a* gives the acceleration at *P*. This suggests

$$\frac{QL}{t^2} \rightarrow \frac{1}{2}a$$

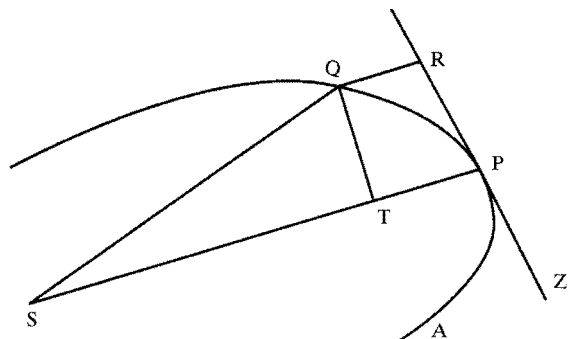
as $t \rightarrow 0$. Sanding top and bottom, Newton could now have shaped the measure QL/t^2 to fit squarely



into his geometric approach. First nudge *L* just a bit along the tangent to the position *R*, making the deviation *QR* exactly parallel to *SP*.

Because time varies as the area in orbital motions, replace *t* by the area of the “sector” *PSQ*, and the sector in turn by the approximating triangle *PSQ*, in the process turning *t* into the product $SP \cdot QT$ — no need to keep tabs on constant factors, such as the missing 1/2 here, for Newton works with proportions, not equations — and the measure QL/t^2 into $QR/(SP \cdot QT)^2$. The limit of this ratio, as $Q \rightarrow P$, gauges the acceleration at *P*. In the *Principia*, this measure of acceleration appears as Corollary I to Proposition VI (Section II, Book I) [11, p. 48]. With this corollary, Newton later derives acceleration laws from orbit shapes.

Cor 1. If a body *P* revolving about the center *S* describes a curved line *APQ*, which a right



line ZPR touches in any point P ; and from any other point Q the curve, QR is drawn parallel to the distance SP , meeting the tangent in R ; and QT is drawn perpendicular to the distance SP ; the centripetal force will be inversely as the solid $SP^2 \cdot QT^2 / QR$, if the solid be taken of that magnitude which it ultimately acquires when the points P and Q coincide.

Before we leave the topic of acceleration, we should take a moment to discuss the role of force and mass in the early sections of the *Principia*. The word ‘force’ appears, as it does above in Corollary I, in many of the definitions, axioms, corollaries, and propositions of the *Principia*, but in the first ten sections, where Newton attends to the one-body problem, force, and mass as well, exist literally in name only, playing no part in the *mathematics*. He may talk of ‘force,’ but Newton calculates accelerations. The Cartesians, Huygens and Leibniz among them, claimed that Newton, by introducing gravity, and therefore action at a distance, brought Aristotelian ‘occult qualities’ back into physics. But he should plead innocent to this charge. In the *Principia*’s work on orbital motions, ‘force’ and ‘gravity’ become merely convenient words, as Newton stresses the relations and laws, with no comment on causes. The cause of gravity comes up only in a General Scholium on the final pages of the *Principia* [II, p. 547]: “But hitherto I have not been able to discover the cause of those properties of gravity from phenomena,” wrote Newton,

and I frame no hypotheses; for whatever is not deduced is to be called an hypothesis; and hypotheses, whether metaphysical or physical, whether of occult qualities or mechanical, have no place in experimental philosophy . . . And to us it is enough that gravity does really exist, and act according to the laws which we have explained, and abundantly serves to account for all the motions of the celestial bodies, and of our sea.

Wouldn’t Newton, that lover of geometry and curvature, have been delighted with Einstein’s view that geometry, indeed the curvature of spacetime, is the very cause of gravity?

After this interlude on Newton’s measure of acceleration, we remain in the past, looking for the original proof of the Acceleration Theorem in the *Principia*.

5

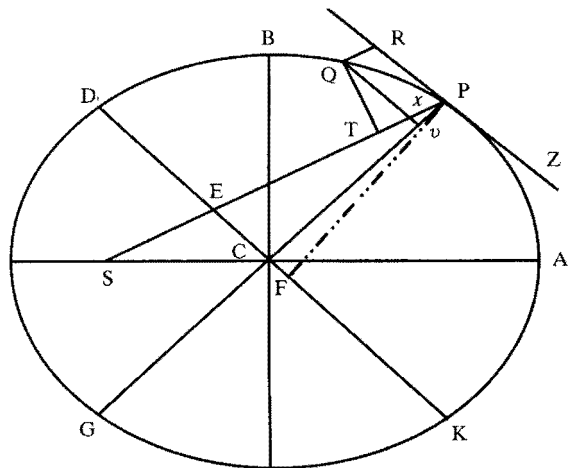
Wasting no time after Corollary I to Proposition VI, Newton attacks a series of problems with his new measure of acceleration. In Propositions VII through XIII, he calculates the acceleration law for circular motions about any given point, semicircular motions about a point infinitely remote, spiral motions about the pole, elliptical motions about the center, and then, in a stately section all their own, elliptical, hyperbolic, and parabolic motions about the focus. Taken together, this final triumphant trio of propositions (XI, XII, and XIII) establishes the Acceleration Theorem: *Every conic motion about the focus has an inverse-square acceleration.*

Newton could have proved the Acceleration Theorem in a single proposition covering general conic motions, but “. . . because of the dignity of the Problem . . .,” he writes, “I shall confirm the . . . cases by particular demonstrations.” [11, p. 57] These “particular demonstrations” naturally offer the same argument with minor variations, so we may safely choose one of the propositions to represent all three. Turn then to the most celebrated page of the *Principia* and to Newton’s analysis for Proposition XI:

Proposition XI Problem VI

If a body revolves in an ellipse; it is required to find the law of the centripetal force tending to the focus of the ellipse.

In the ellipse, Newton draws conjugate diameters DK and PG , with DK parallel to the tangent RPZ . (The midpoints of parallel chords in an ellipse lie on a line, called a *diameter* of the ellipse, and the parallel chords are then called the *ordinates* of the diameter. Two diameters with the property that each



bisects every chord parallel to the other are said to be *conjugate diameters*.) From Q he drops three lines: QR parallel to the focal radius SP , QT perpendicular to SP , and Qx completing the parallelogram $QxPR$. He then extends Qx until it meets PG at v and draws PF perpendicular to DK .

Newton's analysis requires the services of three lemmas, one of his own and two well known to Apollonius of Perga. (For the two Apollonian lemmas, see [1, I p. 15 and VII p. 31] or [18, pp. 151 and 169].)

Newton's Lemma. $PE = AC$

Lemma 1. *All parallelograms circumscribed about any conjugate diameters of an ellipse have equal area.*

Lemma 2. *In an ellipse, the squares of the ordinates of any conjugate diameter are proportional to the rectangles under the segments which they make on the diameter.*

As we have seen in the previous section, Newton measures the acceleration of an orbital motion by computing the limit of the ratio

$$\frac{QR}{(SP \cdot QT)^2}$$

as $Q \rightarrow P$. To infer an inverse-square acceleration for this case of elliptical motion about the focus, he must therefore prove that QR/QT^2 has a limit independent of P . In fact, as we now show, Newton's argument reveals that QT^2/QR tends to the *latus rectum* of the ellipse.

Because QR is Px and (by Newton's Lemma) PE is AC , the similarity of the triangles PxV and PEC implies

$$QR = \frac{Pv \cdot AC}{PC}.$$

On the other hand, Newton's Lemma (again) and the similarity of the triangles QxT and PEF give

$$QT = \frac{Qx \cdot PF}{AC} = \frac{Qx \cdot BC}{CD},$$

where the second equality follows from Lemma 1, which assures us that $PF \cdot CD = BC \cdot AC$. We infer

$$\frac{QT^2}{QR} = \frac{Qx^2 \cdot BC^2}{CD^2} \cdot \frac{PC}{Pv \cdot AC} = \frac{1}{2}L \frac{Qx^2 \cdot PC}{Pv \cdot CD^2},$$

where we have replaced $2BC^2/AC$ by L . (Following Apollonius, Newton calls $2BC^2/AC$ the *latus*

rectum.) If now $Q \rightarrow P$, this last expression has the same limit as

$$\frac{1}{2}L \frac{vG}{PC},$$

for Qv/Qx tends to one and Lemma 2 implies

$$\frac{Qv^2}{Pv \cdot vG} = \frac{CD^2}{PC^2}.$$

But $vG \rightarrow 2PC$, so that $\frac{1}{2}L(vG/PC)$, and thus also QT^2/QR , must tend to L . This completes Newton's analysis for Proposition XI: *Every elliptical motion about the focus has an inverse-square acceleration.*

6

We have been "going under with the swirls and coming out with the eddies, following along the way the water goes," but now just one quick swirl remains: to return from the *Principia* to the present, from Newton's original work on the Acceleration Theorem to the delightful contrast of a contemporary argument.

Any conic motion $\mathbf{r} = \mathbf{r}(t)$ about the focus must satisfy the vector-conic equation (2),

$$\mathbf{r} \cdot (\mathbf{e} + \mathbf{U}) = l,$$

for some positive constant l and constant vector \mathbf{e} . Since \mathbf{r} is an orbital motion, $\mathbf{h} = \mathbf{r} \times \mathbf{v}$ is a constant vector. Since \mathbf{r} is a conic motion,

$$\mathbf{L} = \frac{l}{h^2} \mathbf{v} \times \mathbf{h} - \mathbf{U}$$

is a second constant vector (equal to the eccentricity vector \mathbf{e} by (3)). Differentiating \mathbf{L} yields

$$\mathbf{0} = \frac{l}{h^2} \mathbf{a} \times \mathbf{h} - \frac{\mathbf{h} \times \mathbf{r}}{r^3},$$

and taking lengths we uncover an inverse-square acceleration

$$a = \frac{h^2}{l} \frac{1}{r^2},$$

proving again

Newton's Acceleration Theorem. *Every conic motion about the focus has an inverse-square acceleration.*

References

1. Apollonius of Perga, *Treatise on Conic Sections*, Volumes I–VII, Cambridge University Press, Cambridge, 1896.

2. J. L. Axtell, Locke, Newton, and the Two Cultures, *John Locke: Problems and Perspectives*, Cambridge University Press, Cambridge, 1969, 165–182.
3. J. B. Brackenridge, Newton's Unpublished Dynamical Principles: A Study in Simplicity, *Annals of Science* 47 (1990), 3–31.
4. —, The Critical Role of Curvature in Newton's Developing Dynamics, *The Investigation of Difficult Things: Essays on Newton and the History of the Exact Sciences*, edited by P. M. Harman and A. E. Shapiro, Cambridge University Press, Cambridge, 1992, 231–260.
5. —, *The Key to Newton's Dynamics: The Kepler Problem and the Principia*, University of California Press, Berkeley, 1995.
6. S. Chandrasekhar, *Newton's Principia for the Common Reader*, Oxford University Press, New York, 1995.
7. G. E. Christianson, *In the Presence of the Creator: Isaac Newton and His Times*, The Free Press, New York, 1984.
8. I. B. Cohen, *Introduction to Newton's 'Principia'*, Harvard University Press, Cambridge, 1971.
9. N. Grossman, *The Sheer Joy of Celestial Mechanics*, Birkhäuser, New York, 1995.
10. M. Nauenberg, Newton's Early Computational Method for Dynamics, *Archive for History of Exact Sciences* 46 (1994), 221–252.
11. I. Newton, *Sir Isaac Newton's Mathematical Principles of Natural Philosophy and His System of the World*, original translation by A. Motte in 1729, revised by F. Cajori, University of California Press, Berkeley, 1946.
12. —, *The Mathematical Papers of Isaac Newton*, Volumes I–VIII, edited by D. T. Whiteside, Cambridge University Press, Cambridge, 1967–1981.
13. —, *The Correspondence of Isaac Newton*, edited by A. R. Hall and L. Tilling, Cambridge University Press, Cambridge, 1975.
14. —, *The Preliminary Manuscripts for Isaac Newton's 1687 Principia 1684–1685*, introduction by D. T. Whiteside, Cambridge University Press, Cambridge, 1989.
15. B. Pourciau, On Newton's Proof That Inverse-Square Orbits Must be Conics, *Annals of Science* 48 (1991), 159–172.
16. —, Newton's Solution of the One-Body Problem, *Archive for History of Exact Sciences* 44 (1992), 125–146.
17. —, Radical Principia, *Archive for History of Exact Sciences* 44 (1992), 331–363.
18. G. Salmon, *A Treatise on Conic Sections*, Chelsea Publishing Company, New York, 1954.
19. Chuang Tzu, *Chuang Tzu: Basic Writings*, translated by Burton Watson, Columbia University Press, New York, 1964.
20. R. Weinstock, Isaac Newton: Credit Where Credit Won't Do, *College Mathematics Journal* 25 (1994), 179–193.
21. R. Westfall, *Never at Rest: A Biography of Isaac Newton*, Cambridge University Press, Cambridge 1980.
22. D. T. Whiteside, The Prehistory of the Principia From 1664 to 1686, *Notes and Records of the Royal Society of London* 45 (1991), 11–61.
23. C. Wilson, Newton's Orbit Problem: A Historian's Response, *College Mathematics Journal* 25 (1994) 193–201.

Newton as an Originator of Polar Coordinates

C. B. BOYER

American Mathematical Monthly 56 (1949), 73–78

The name of Newton, indissolubly linked with the calculus, seldom is associated with analytic geometry, a field to which he nevertheless made important contributions. Newton's use of polar coordinates, for example, seems to have been overlooked completely in the historiography of mathematics. The polar coordinate system is ascribed generally [1] to Jacques Bernoulli in 1691 and 1694, although it has been attributed [2] to others as late as Fontana in 1784. It is the purpose here to call attention to an application of polar coordinates made by Newton probably a score of years before the earliest publication of Bernoulli's work.

In the Horsley edition of the *Opera* of Newton there appears a treatise entitled *Artis analyticae specimina vel Geometria analytica* [3] which is essentially the same as the Newtonian *Method of fluxions*, published in 1736 by Colson. The discrepancy in titles — *Geometria analytica* or *Method of fluxions* — conveniently indicates that the work treats of coordinate geometry as well as the calculus. In fact, its analytic form stands in marked contrast to the synthetic style of the *Principia*, which also contained some elements of the calculus. The *Method of fluxions* makes systematic use of coordinates in problems on tangents, curvature, and rectification. Moreover, Newton did not limit himself, as had his predecessors, to a single type of coordinate system. Having shown how to use fluxions in finding tangents to curves given in terms of Cartesian coordinates, oblique as well as rectangular, Newton included some examples illustrating other types. In connection with these he gave, informally, the equivalent of equations of transformation for polar and rectangular coordinates, $xx + yy = tt$ and $tv = y$, where t is the radius vector and v is a line representing the sine of the vectorial angle associated with

the point (x, y) . Following these exercises, Newton proceeded to give a more definitive account of non-Cartesian systems:

However it may not be foreign from the purpose, if I also shew how the problem may be perform'd, when the curves are refer'd to right lines, after any other manner whatever; so that having the choice of several methods, the easiest and most simple may always be used [4].

To illustrate his point, Newton suggested eight new types of coordinate system, made up of various combinations of pairs of distances measured radially from given points, or obliquely to given fixed lines, or curvilinearly along arcs of circles. One of the new systems — Newton refers to it as the “Seventh Manner; For Spirals” — is essentially that now known as polar coordinates.

Let A be the center and AB a radius of the circle BG (Figure 1), and let D be any point on the curve

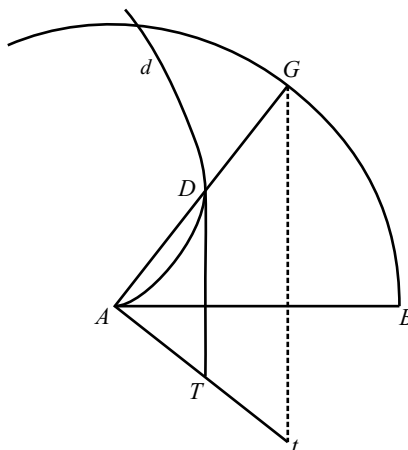


Figure 1.

ADd. Then, designating *BG* by *x* and *AD* by *y*, the curve *ADd* is determined by a relationship between *x* and *y*. Newton suggested

$$x^3 - ax^2 + axy - y^3 = 0$$

as an illustration, and for this curve he determined, from the proportion

$$\dot{y} : \dot{x} :: AD : At,$$

the polar subtangent *AT* for a point *D*. Similarly Newton found the polar subtangents of $y = ax/b$, “which is the equation to the spiral of Archimedes”, and the curve $by = xx$; and, he concluded, “thus tangents may be easily drawn to any spirals whatever” [5].

Following the calculation of the radius of curvature for rectangular Cartesian coordinates *x* and *y*,

$$r = \frac{1 + zz}{\dot{z}} \sqrt{1 + \dot{z}z/\dot{z}}$$

where $z = \dot{y}$ and fluxions of independent variables are taken as unity, Newton again turned to the corresponding problem in polar coordinates. Using a diagram and a notation similar to those applied in connection with tangent problems — but with the radius *AB* of the reference circle taken as unity — he derived the result

$$r \sin \psi = \frac{y + yzz}{1 + zz - \dot{z}},$$

where $z = \dot{y}/y$ and ψ is the angle between the tangent and the radius vector (fluxions of independent variables again being taken as unity). Newton applied this formula, virtually the same as the modern equivalent, to the spiral of Archimedes and to the curves $ax^2 = y^3$ and $ax^2 - xy = y^3$. In conclusion he added, “And thus you will easily determine the curvature of any other spirals; or invent rules for any other kinds of curves”. That he realized the significance of his use of polar coordinates seems to be implied by his further comment that he here had “made use of a method which is pretty different from the common ways of operation” [6].

The comparison of the parabola with the spiral was a favorite topic of the seventeenth century, and in his treatment of this question, in the *Method of fluxions*, Newton made use of a polar coordinate system yet a third time. Here, however, his scheme differed from that previously presented. The notation, too, was modified, but this may have been done in order to avoid confusion in the simultaneous use of

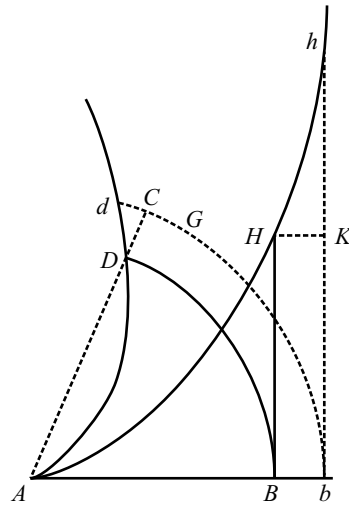


Figure 2.

polar and Cartesian coordinates. If *D* is any point on a curve *ADd*, Newton took the coordinates of *D* as *z* and *v*, where *z* is the radius vector *AD* and *v* is the circular arc *BD* (Figure 2). That is, whereas his earlier coordinates were, in modern notation, $(r, a\theta)$, Newton this time used $(r, r\theta)$. Then if the relation between *z* and *v* is given “by means of any equation”; and if a new curve *AHh*, given in rectangular coordinates $AB = z$ and $BH = y$, is so determined that, for all corresponding positions of *D* and *H*, the arc *AD* is equal to the arc *AH*; then Newton showed that

$$\dot{y} = \dot{v} - v\dot{z}/z,$$

or, if \dot{z} is taken as unity, $\dot{y} = \dot{v} - v/z$. In particular, “if $zz/a = v$ is given as the spiral of Archimedes”, then $\dot{v} = 2z/a$, and hence $z/a = \dot{y}$ and $zz/(2a) = y$. The lengths of the spirals $z^3 = av^2$ and $z\sqrt{a+z} = v\sqrt{c}$ are shown in like manner [7] to correspond respectively to lengths measured along the semi-cubical parabola $z^{3/2} = 3a^{1/2}y$ and the curve

$$(z - 2a)\sqrt{ac + cz} = 3cy.$$

Evidence indicates [8] that the *Method of fluxions* was composed by 1671, at which time Jacques Bernoulli was in his teens; and there seems to be no reason for suspecting the sections on polar coordinates as later interpolations. The three pertinent passages would appear to be a natural part of the whole; and Horsley, after his editorial examination of three different manuscript copies of the work, apparently saw no reason to question the date or authenticity of this material. It is surprising therefore

that this contribution to coordinate geometry should have gone unnoticed so completely that the use of polar coordinates invariably is attributed to others of later periods. Newton is not entitled to priority of publication, for the work appeared posthumously in 1736; but evidence indicates that he was the first one to adopt a system of polar coordinates in strictly analytic form [9]. Moreover, his work in this connection is superior, in flexibility and generality, to any similar proposal to appear during his lifetime.

Priority in the publication of polar coordinates seems to go to Jacques Bernoulli who in the *Acta Eruditorum* of 1691 proposed measuring abscissas along the arc of a fixed circle and ordinates radially along the normals. Three years later, however, he presented in the same journal a system identical, both in conception and notation, with that first proposed by Newton. He used the coordinates y and x , where y is the length of the radius vector of the point and x is the arc cut off by the sides of the vectorial angle on a circle of radius a described about the pole as center. That is, Bernoulli too adopted the coordinates $(r, a\theta)$, whereas in his earlier work he had used a less convenient system equivalent to $(a - r, a\theta)$. Bernoulli, like Newton, was interested primarily in applications of his system to the calculus; and so he also derived a formula for radius of curvature in polar coordinates [10] and applied it to the spiral of Archimedes, $y = ax : c$.

The polar coordinates of Newton and Bernoulli in 1704 were applied by Varignon [11] to a comparison of the higher parabolas and spirals of Fermat, but no reference was made to Newton's work. Varignon ascribed the idea to Jean Bernoulli and gave to Jacques Bernoulli only the credit for priority of publication. His information in this connection was perhaps not unbiased; and his treatise is tedious and unimaginative in comparison with the work of Newton, at that time still unpublished.

In 1729, two years after Newton's death, Hermann approached polar coordinates from a new point of view. He did not concern himself with spirals, as had Newton, Bernoulli, and Varignon, nor was he chiefly interested in the calculus. He proposed the study of loci "through the relationship which vectorial radii bear to the sine or cosine of the angles of projection, from the consideration of which the properties of curves flow just as elegantly as they are brought out in the usual manner" [12]. That is, Hermann seems first to have thought of polar coordinates as a part of analytic geometry proper. He gave equations for transforming from Cartesian to

polar coordinates, and he applied his new system to a number of algebraic curves, including the conics. It should be noted, however, that he did not express his equations specifically in modern form, but wrote them in terms of z , m , and n , where z is the radius vector and m and n are the sine and cosine respectively of the vectorial angle. Moreover, where his predecessors had applied the polar system to spirals alone, Hermann inversely used the scheme exclusively for algebraic curves.

Euler in 1748 seems to have been the first one to combine the points of view of Newton and Hermann. In the influential *Introductio in analysin infinitorum* he devoted a large portion of each of two chapters to polar coordinates, one dealing with algebraic curves and the other with spirals. In the first case [13] he gave the equations of transformation

$$x = z \cos \phi, \quad y = z \sin \phi,$$

introducing modern trigonometric symbolism into polar coordinates. He gave general consideration to z as a function of $\sin \phi$ and $\cos \phi$, and he noted in more detail the limaçons

$$z = b \cos \phi \pm c$$

and the conchoids

$$z = \frac{b}{\cos \phi \pm c}.$$

In the treatment of transcendental curves Euler adopted a slightly different notion and notation for the independent variable in polar coordinates [14]. Here he studied curves of the form $z = f(s)$, where the argument s is the arc of a unit circle which measures the angle ϕ , feeling, apparently, that coordinates must of necessity denote lengths. In connection with the spiral curves which he drew, Euler made use of the general angle, allowing s to increase indefinitely, both positively and negatively. The spiral of Archimedes therefore appeared, perhaps for the first time, in its dual form [15]. The work of Euler is so thorough and systematic that polar coordinates frequently are attributed to him [16]. Certainly no one after him deserves credit as the inventor of the system. Fontana in 1784 did perhaps supply the name "polar equation" of a curve [17], and he may have been first [18] in studying analytically curves of the form $r = f(\theta, \sin \theta, \cos \theta)$; but one gets the very definite impression that his ideas and manner of treatment were inspired by Euler. It is probably not too much to say that although Newton probably

originated polar coordinates, it was the work of Euler which was the decisive factor in making the system a traditional part of elementary analytic geometry. Polar coordinates gradually achieved greater prominence until in 1857 there appeared an entire volume devoted to the analytic geometry of this system in the plane and in space [19]. In 1874 the system was generalized to include elliptic polar coordinates and hyperbolic polar coordinates [20].

It may not be inappropriate to point out here that bipolar coordinates, recently ascribed [21] to Cournot in 1847, also were proposed by Newton. Such a system appeared in the *Method of fluxions* as the “Third Manner” of determining a curve. Here Newton considered [22] the “ellipses of the second order”, now known as “ovals of Descartes”. In *La géométrie* [23] Descartes had proposed these curves in connection with problems in refraction, but he handled them, as Newton remarked, “in a very prolix manner”, without the application of coordinates. Newton therefore seems to have been the originator of bipolar coordinates in the strict sense. Representing by x and y the “subtenses” (or distances) of a variable point from two fixed points (or poles), Newton wrote “their relation” for the ovals as

$$a + ex/d - y = 0.$$

From this equation he found the ratio of the fluxions, and hence the tangent line. Newton pointed out further that for $a - ex/d - y = 0$, a contrary sense is indicated in the construction; and he noted that if $d = e$, the curve becomes a conic section. He closes this topic with the remark that “it would be easy . . . to give more Examples”.

Newton’s generalizations of the coordinate idea may not be among his greatest contributions to mathematics, but they do entitle him to a larger place in the history of analytic geometry. In this field, as well as in infinitesimal analysis, one may appropriately declare, *Ex ungue leonem*.

References

1. See, e.g., J. L. Coolidge, *A history of geometrical methods* (Oxford, 1940), p. 111. Cf. Florian Cajori, *History of mathematics* 2nd ed., New York, 1931, pp. 221, 224.
2. See, e.g., D. E. Smith, *History of mathematics* (2 vols., New York, c. 1923-1925), II, 324. Cf. Moritz Cantor, *Vorlesungen über Geschichte der Mathematik* (4 vols., Leipzig, 1880-1908), IV, 513.
3. *Opera quae exstant omnia* (5 vols., Londini, 1779-1785), I, 389-518. This appears also (with another title) in Newton’s *Opuscula* (3 vols., Lausannae, 1744), I, 29-200.
4. Sir Isaac Newton, *The method of fluxions and infinite series* (transl. by John Colson, London, 1736), p. 51. Cf. *Opera*, I, 435.
5. *Method of fluxions*, p. 56, *Opera*, I, 440.
6. *Method of fluxions*, p. 70, *Opera*, I, 453.
7. *Method of fluxions*, pp. 132-134, *Opera*, I, 511-512.
8. See, e.g., H. G. Zeuthen, *Geschichte der Mathematik im XVI. und XVII. Jahrhundert* (Leipzig, 1903), p. 374; and Coolidge, *op. cit.*, p. 320. Cf. also Newton, *La méthode des fluxions* (Paris, 1740), Buffon’s Preface.
9. More or less vague adumbrations of the idea of polar coordinates can, of course, be found in earlier work going back as far as the time of Archimedes’ spiral. In early works on perspective the use of concentric circles and radiating lines in problems relating to “deformations” or “anamorphoses” represents a non-analytic application of the polar coordinate idea, much as ideas of latitude and longitude were forerunners of Cartesian coordinates. Some work of James Gregory in 1668 also represents to some extent an anticipation of such a system. See James Gregory, *Tercentenary memorial volume* (ed. by H. W. Turnbull, London, 1939), p. 493.
10. See Jacques Bernoulli, *Opera* (2 vols., Genevae, 1744), I, 432, 578-580. Cf. also a note by Eneström in *Bibliotheca Mathematica* (3), XIII (1912-1913), 76-77.
11. Pierre Varignon, Nouvelle formation de spirales, *Académie des Sciences, Mémoires*, 1704, pp. 69-131. Cf. *Académie des Sciences, Histoire*, 1704, pp. 47-57. Generalized spirals in polar coordinates appeared also in C. R. Reyneau, *Usage de l’analyse* (vol. II, Paris, 1708), p. 593 and in J. B. Caraccioli, *De lineis curvis* (Pisis, 1740), p. 161.
12. Jacob Hermann, Consideratio curvarum in punctum positione datum projectarum, et de affectionibus earum inde pendentibus, *Commentarii Academiae Petropolitanae*, IV (1729), 37-46.
13. Leonhard Euler, *Introductio in analysin infinitorum* (2 vols., Lausannae, 1748), II, 212 ff.
14. *Ibid.*, II, 284 ff.
15. Gino Loria, Perfectionnements, évolution, métamorphoses du concept de ‘coordonnées’, *Mathematica*, XVIII (1942), 125-145; XX (1944), 1-22, incorrectly ascribes this dual form to Cournot a century later.
16. See, e.g., E. Müller, Die verschiedenen Koordinatensysteme, *Encyklopädie der mathematischen Wissenschaften*, III (1), 596-770, especially, pp. 656-657.

- Cf. *Encyclopédie des sciences mathématiques*, III (3), 1, p. 47. See also Loria, *loc. cit.*
17. Gregorio Fontana, *Sopra l'equazione d'una curva*, *Memorie di matematica e fisica della società italiana*, II (part 1, 1784), 123–141. See especially p. 128.
 18. Gregorio Fontana, *Disquisitiones physico-mathematicae* (Papiae, 1780), pp. 184–185. Fontana used y instead of r , and for θ he took x , where x is the arc of a unit circle measuring the angle θ .
 19. J. A. Grunert, *Analytische Geometrie der Ebene und des Raumes für polare Koordinaten* (Greifswald and Leipzig, 1857).
 20. C. A. Laisant, *Essai sur les fonctions hyperboliques* (Paris, 1874), pp. 71–83.
 21. Loria, *op. cit.*, p. 138. Loria here overlooks also the use of bipolar coordinates by L. N. M. Carnot, *Géométrie de position*, (Paris, 1803), p. 469, and J. D. Gergonne, *Annales de mathématiques*, IV, 1813–14, p. 42 f.
 22. *Method of fluxions*, p. 54 f, *Opera*, I, 437 f.
 23. See *The Geometry of Descartes* (translated by D. E. Smith and M. L. Latham, Chicago and London, 1925), p. 114 ff.

Newton's Method for Resolving Affected Equations

CHRIS CHRISTENSEN

College Mathematics Journal 27 (1996), 330–340

During the 300 years since Newton and Leibniz began disputing which of them had discovered the calculus, debates have continued over the credit due to Newton for various scientific and mathematical achievements. Recent research by Nick Kollerstrom [11] has led him to credit Thomas Simpson (1710–1761) with the first discovery and publication in 1740 [18] of what is now called Newton's method. William Dunham [8] has pointed out the irony that Newton, who "bitterly resented people's getting credit for results they did not *originally* discover," is credited with a method of approximation that "in its full generality seems to be due to" Simpson.

The debate over priority for Newton's method may now be settled, but almost forgotten in the discussion is that Newton presented his method for approximating real roots side by side with a similar method for writing y in terms of x when y is implicitly defined in terms of x by a polynomial equation—a so-called "affected equation." This second "Newton's method" is an important tool in modern algebraic geometry and, although it is more subtle than his method for approximating roots, it can be understood by precalculus students.

Richard S. Westfall [22] highlights how Newton used his method for resolving affected equations to integrate algebraic equations:

... in the mid-1660s, Newton was working toward a general method of squaring curves, as they put it then; let us say "integration" for simplicity. Earlier mathematicians on whom he drew had established the algorithm for integrating simple powers and had understood that polynomials can be integrated by the same procedure term by term. What was one to do

with curves such as $y = (1 - x^2)^{1/2}$ and $y = 1/(1 - x)$? With the binomial theorem Newton succeeded in expanding such curves into infinite series that he could integrate term by term and thus approximate the answer for some value of x however close he chose. Later he developed an iterative procedure by which to expand "affected" equations ... into infinite power series in [fractional] powers of x . With that he had a general method to integrate all of the algebraic equations then known to mathematics. No earlier mathematician had even approached a method of this power.

Though Newton's disciples Stirling and Taylor fully appreciated his second method, it "seems to have been lost sight of ... after their time" [7, p. 396]. S. Abhyankar [2] has sketched its subsequent history:

Newton's theorem was revived by Puiseux in 1850 [16], so it acquired the name "Puiseux expansion" which is a misnomer. What's more is that Puiseux's proof, being based upon Cauchy's integral theorem, applies only to convergent power series with complex coefficients. On the other hand, Newton's proof, being algorithmic, applies equally well to power series, whether they converge or not. Moreover, and that is the main point, Newton's algorithmic proof leads to numerous other existence theorems while Puiseux's existential proof does not do so.

In what follows, I will first briefly examine Newton's method for approximating real roots and compare it to the method of Raphson and the one found in today's calculus texts. Then I will show how New-

ton generalized his method for approximating real roots to a method for resolving affected equations.

Both of Newton's methods appear in his *Methods of series and fluxions* [13] (composed in 1671 but first published in 1736), in his *On analysis by infinite equations* [12] (1669), and in his two famous letters to Leibniz in 1676 — the *Epistola prior* [14] and the *Epistola posterior* [15]. In each, Newton gives, as examples, approximating a real root of the polynomial equation $y^3 - 2y - 5 = 0$ and resolving the affected equation $y^3 + axy + a^2y - x^3 - 2a^3 = 0$. I will refer to Newton's more compact exposition of the methods in the two letters, relevant portions of which are included here as Appendices I and II.

Leibniz, in 1674, wrote to Oldenburg, secretary of the Royal Society, saying that he possessed "general analytic methods depending on power series." Oldenburg in reply told him that Newton and Gregory had used such series in their work. In answer to a request for information, Newton wrote the *Epistola prior* on June 13, 1676, giving a brief account of his method. He here enunciated the binomial theorem along with his methods for approximating real roots and resolving affected equations. Leibniz replied on August 27 asking for fuller details, and Newton sent through Oldenburg an account of the way in which he had been led to some of his results, the *Epistola posterior* of October 24, 1676.

Newton's method for approximating roots

Newton, in the *Epistola prior* (see the first table in Appendix I), exhibits his method of approximating real roots by way of the cubic $y^3 - 2y - 5 = 0$. He first guesses that there is a root near $y = 2$. But $y = 2$ is not a solution to the equation; so Newton modifies his guess slightly by substituting $y = 2 + p$ into the equation and obtains

$$p^3 + 6p^2 + 10p - 1 = 0. \quad (1)$$

He neglects the non-linear terms "on account of their smallness", so the linear portion of the polynomial must vanish; that is, $10p + 1 = 0$. Thus $p = 1/10$, and Newton's new approximation for the root is $y = 2 + p = 2 + 0.1$.

Because $y = 2.1$ is not a root, he repeats the process by modifying p . He substitutes $p = 0.1 + q$ into the equation in p , which yields

$$q^3 + 6.3q^2 + 11.23q + 0.061 = 0. \quad (2)$$

Again neglecting higher-order terms, he selects q so that $11.23q + 0.061 = 0$. Therefore, $q = -0.061/11.23 = -0.0054$, and Newton's approximation for the root is now $y = 2 + 0.1 - 0.0054$.

Similarly, q is modified. He substitutes $q = -0.0054 + r$ into the equation in q , getting

$$r^3 + 6.2838r^2 + 11.162r + 0.000541551 = 0, \quad (3)$$

and he selects r so that $11.162r + 0.0005416 = 0$. So,

$$r = -0.0005416/11.162 = -0.00004852.$$

The example ends with $y = 2 + 0.1 - 0.0054 - 0.00004852$ as an approximation to the real root near $y = 2$.

Raphson's method applied to Newton's cubic.

What if we apply Raphson's method to Newton's cubic? We might begin, as Newton did, by guessing that there is a root near $y = 2$, substituting $y = 2 + p$ into the cubic to obtain (1). Like Newton, we would let $10p - 1 = 0$, solve to get $p = 1/10$, and obtain $y = 2 + 0.1$ as the new approximation.

Here is where their methods differ. Newton lets $p = 0.1 + q$ and substitutes into *the equation in terms of p*, whereas Raphson would let $y = 2.1 + q$ and would substitute this into *the original cubic*. Raphson too obtains (2), which he would solve by the same reasoning as Newton, making his next approximation for the root be $y = 2.1 + q = 2.0946$.

If Raphson continued, he would let $y = 2.0946 + r$ and substitute this, again, into *the original cubic*, where Newton substituted $q = -0.0054 + r$ into (2), *the expression for the original cubic in terms of the variable q*. Although the methods are algebraically equivalent, essentially Newton expresses his algorithm *recursively* where Raphson expresses it as a simple *iterative* procedure.

Cajori [6] compares Newton's method and Raphson's method:

In 1690, Joseph Raphson (1648–1715), a Fellow of the Royal Society of London, published a tract, *Analysis aequationum universalis* [17]. His method closely resembles that of Newton. The only difference is this, that Newton derives each successive step p , q , r , of approach to the root, from a new equation, while Raphson finds it each time by substitution in the original equation. . . . Raphson does not mention Newton; he evidently considered the difference sufficient for his method to be classed independently.

By returning at each step to the original polynomial, Raphson expresses the procedure as iteration of a function $x_{n+1} = g(x_n)$ where $g(x) = x - f(x)/f'(x)$ (although he does not identify the denominator as a derivative). To see the familiar form, consider the general cubic equation

$$f(y) = c_0y^3 + c_1y^2 + c_2y + c_3 = 0.$$

Let $y = x + p$ where x is an initial guess for a root, and expand:

$$\begin{aligned} c_0(x+p)^3 + c_1(x+p)^2 + c_2(x+p) + c_3 &= 0 \\ c_0(x^3 + 3x^2p + 3xp^2 + p^3) + c_1(x^2 + 2xp + p^2) \\ &+ c_2(x+p) + c_3 = 0 \\ c_0p^3 + (3c_0x + c_1)p^2 + (3c_0x^2 + 2c_1x + c_2) \\ &+ (c_0x^3 + c_1x^2 + c_2x + c_3) = 0. \end{aligned}$$

Then discard all terms in p of degree greater than 1 and solve for p :

$$p = -\frac{c_0x^3 + c_1x^2 + c_2x + c_3}{3c_0x^2 + 2c_1x + c_2} = -\frac{f(x)}{f'(x)}.$$

Raphson's approach may be conceptually simpler, but note that Newton's recursive arrangement is ideal for hand calculations. Also, the quadratic convergence of the method is clear from Newton's calculation—the approximate doubling of the number of correct digits with each iteration. Newton recognized this [13, p. 47], for at each stage of the calculation he omitted all terms whose contribution would affect only the insignificant digits.

Kollerstrom [11] states that “What is *today* known as ‘Newton’s method of approximation’ has two vital characteristics: it is *iterative*, and it employs differentials.” He argues that because Newton did not return to the original equation for his substitutions his method fails to be iterative, and “it did not employ any fluxional [differential] calculus.” Raphson's method, though iterative, likewise made no use of differential calculus. Only Simpson's version “sufficiently resembles the modern formulation for him to be credited . . . as inventor,” Kollerstrom concludes.

While not contesting the accuracy of Kollerstrom's analysis, I leave it to the reader to decide how much credit Newton deserves for this method of approximating roots.

Resolution of affected equations

Recall that one of Newton's goals was to be able to integrate y when this variable is implicitly defined

as a function of x by a polynomial equation in x and y . Once Newton was able to expand y in terms of a (fractional) power series in terms of x , he could integrate term by term. As we shall see, the key to the method of expansion he used is a geometric device, the *Newton polygon*. Otherwise, this method is very similar to Newton's recurrence method for approximating roots.

Newton begins with the affected equation

$$y^3 + axy + a^2y - x^3 - 2a^3 = 0,$$

containing a parameter a (see Appendix I, *Epistola prior*). He wishes to expand y as a series in powers of x . (Just as for Taylor series, the translation $x \rightarrow x - x_0$ could be used to translate any point to 0, so Newton's procedure could be used to expand y in powers of $x - x_0$.)

In the *Epistola posterior* (see Appendix II), Newton observes that if $x = 0$ then $y^3 + a^2y - 2a^3 = 0$; “hence $y = a$ very nearly.” But as $y = a$ is not a solution of the original equation, it must be modified. Newton modifies the root by adding p to it and substitutes $y = a + p$ into the equation to obtain

$$p^3 + 3ap^2 + 4a^2p + axp + a^2x - x^3 = 0. \quad (4)$$

(See the second table in the *Epistola prior*. Notice that this table looks like the table that Newton obtained when he approximated a real root of $y^3 - 2y - 5 = 0$.)

Following his method for approximating real roots, the next thing to be done would be to set equal to zero the low degree terms of the polynomial in p and x . But which ones should be considered the “low degree terms”? Newton assigned a certain weight to p and determined the total degrees of the terms of the polynomial relative to this weight. He used the Newton polygon to determine the weight assigned to the dependent variable p .

To form this polygon, for each monomial $cx^j p^k$ present in the polynomial we plot the point in the plane with cartesian coordinates (j, k) . For example, Figure 1 shows the points corresponding to the terms of (4). (In Appendix II Newton plots the exponent of x on the vertical axis, however.) We “apply a ruler” (see Appendix II) to determine (together with the half-lines on the x - and y -axes) a convex polygonal path enclosing the points corresponding to the terms of the polynomial. For equation (4), the polygon consists of the line segment joining $(0, 1)$ and $(1, 0)$, the half-line from $(0, 1)$ to $(0, \infty)$, and the half-line from $(1, 0)$ to $(\infty, 0)$.

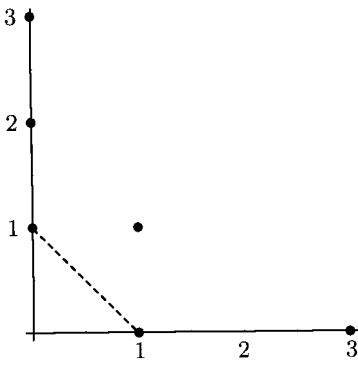


Figure 1. Newton polygon for equation (4)

How does this Newton polygon help find p ? Newton says “I pick out the terms of the equation distinguished by the parallelograms in contact with the ruler, and thence get the quantity to be added to the quotient.” In algebraic form the procedure can be described as follows: The line joining $(0, 1)$ and $(1, 0)$ has slope -1 , so let $-1/\mu = -1$ and solve to obtain $\mu = 1$. Then p is given weight $\mu = 1$ by letting $p = tx^\mu = tx$, where t is a constant to be determined. Substituting $p = tx$ into (4) yields

$$(t^3 - 1)x^3 + (3at^2 + at)x^2 + (4a^2t + a^2)x = 0.$$

Newton, as in his root approximation method, ignores all but the lowest-degree terms in x . Thus he finds a value of t for which $(4a^2t + a^2)x = 0$, namely $t = -\frac{1}{4}$. (Notice that the terms of lowest degree come from $4a^2p$ and a^2x , the terms in (4) that correspond to the corner points on the Newton polygon.) So $p = -\frac{1}{4}x$, and Newton’s new approximation for y is $y = a - \frac{1}{4}x$.

Now substitution shows that $y = a - \frac{1}{4}x$ is not a solution to the original equation; therefore p must be modified. Newton substitutes $p = -\frac{1}{4}x + q$ into (4), the equation in terms of p and x , and obtains

$$q^3 + \left(\frac{3}{4}x + 3a\right) + \left(\frac{3}{16}x^2 - \frac{a}{2}x + 4a^2\right)q - \frac{65}{64}x^3 - \frac{a}{16}x^2 = 0. \quad (5)$$

He then plots the points corresponding to its terms and forms the Newton polygon in Figure 2.

The line segment joining $(0, 1)$ and $(2, 0)$ is an edge of the Newton polygon and has slope $-\frac{1}{2}$. Solving $-\frac{1}{\mu} = -\frac{1}{2}$, gives $\mu = 2$, so q is given weight $\mu = 2$. If we substitute $q = tx^\mu = tx^2$ into (5), the polynomial in terms of q and x , the terms of lowest degree will be those corresponding to the points

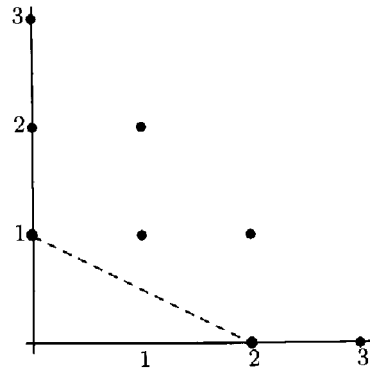


Figure 2. Newton polygon for equation (5)

$(0, 1)$ and $(2, 0)$: $4a^2q$ and $-\frac{a}{16}x^2$. Now we need only substitute $q = tx^2$ into $4a^2q - \frac{a}{16}x^2$, set the result equal to 0, and solve for t :

$$4a^2tx^2 - \frac{a}{16}x^2 = 0, \text{ whence } t = \frac{1}{64a}.$$

Newton’s new approximation is $y = a - \frac{1}{4}x + \frac{1}{64a}x^2$.

We now understand the final line in Appendix II. The term in $y = a + p + q$ coming from q results “from dividing the terms involving the lowest power of the variable x [in (5)] by the coefficient of the root” q .

Again, $y = a - \frac{1}{4}x + \frac{1}{64a}x^2$ is not a solution to the original equation; therefore q must be modified. Newton substitutes $q = \frac{1}{64a}x^2 + r$ into the equation in terms of q and x and obtains

$$r^3 + \frac{3}{64a}x^2r^2 - \frac{3}{4}xr^2 + 3ar^2 + \frac{3}{4096a^2}x^4r - \frac{3}{128a}x^3r + \frac{9}{32}x^2r - \frac{a}{2}xr + 4a^2r + \frac{1}{262144a^3}x^6 - \frac{3}{16384a^2}x^5 + \frac{15}{4096a}x^4 - \frac{131}{128}x^3 = 0. \quad (6)$$

(I resorted to using *Mathematica*!) Figure 3 shows the Newton polygon for this equation.

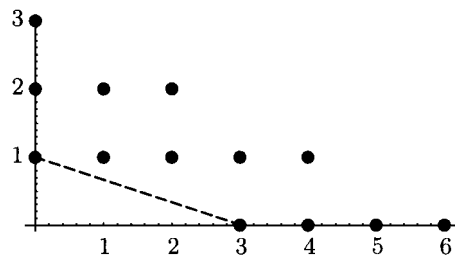


Figure 3. Newton polygon for equation (6)

The line segment joining $(0, 1)$ and $(3, 0)$ has slope $-\frac{1}{3}$. So x is given weight 1 and r gets weight 3. Substituting $r = tx^\mu = tx^3$ into the polynomial in terms of r and x , we find the terms of lowest degree correspond to the points $(0, 1)$ and $(3, 0)$. That is, $4a^2tx^3 - \frac{131}{128}x^3 = 0$, so $t = \frac{131}{512a^2}$. Hence the next approximation is

$$y = a - \frac{1}{4}x + \frac{1}{64a}x^2 + \frac{131}{512a^2}x^3.$$

The process can be continued to find as many terms of a power series expansion of y in terms of x as desired. We can see in Appendix I how efficiently Newton arranges the calculation and how closely it parallels his earlier approximation of a root of his cubic equation.

Fractional exponents are needed. Although Newton's cubic example is a good illustration of the algorithm, it does not exhibit a key feature of the series expansions, namely, the need for fractional exponents. The polynomial curve $y^2 - x^3 = 0$ provides an easy example of this. There are only two points to plot to find the Newton polygon, $(0, 2)$ and $(3, 0)$. The slope of the line segment joining these points is $-\frac{2}{3}$. So x has weight 1 and y has weight $\mu = \frac{3}{2}$. To make

$$(tx^{3/2})^2 - x^3 = (t^2 - 1)x^3$$

equal to 0 we must have $t = \pm 1$, so $y = \pm x^{3/2}$. Substituting, both $y = x^{3/2}$ and $y = -x^{3/2}$ satisfy the equation $y^2 - x^3 = 0$, so the algorithm terminates. The functions $y = x^{3/2}$ and $y = -x^{3/2}$ are the two branches of the curve near the origin, where it has a cusp; see Figure 4.

More generally, the various power series expansions obtained by using the Newton polygon correspond to the branches of a polynomial curve. That is why algebraic geometers use these expansions to

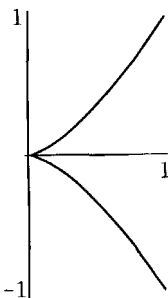


Figure 4. The curve $y^2 - x^3 = 0$

describe the curve near a point. For example, graphing polynomial approximations of the power series expansions of the branches is one way to plot the graph of y as a function of x near a point.

Extensions

For Newton's cubic the polygon has only a single line segment (other than the half-lines on the axes). For "irreducible" curves, the polygon always has only one line segment. (The converse, however, is not true [3, pp. 185, 186].) If the Newton polygon has more than one line segment, we choose the steepest negative slope (as Newton indicates in the *Epistola posterior*).

Surprisingly, the denominators of the exponents do not increase indefinitely; there is a positive integer m that suffices for all denominators. Therefore, y can be expressed as a power series in t (with integer exponents) where $t = x^{1/m}$.

Over a century ago Halphen [9] and Smith [19] pointed out that a certain finite number of terms of the series expansion of a branch are crucial. These terms determine a finite set of pairs of positive integers called the *characteristic (Puiseux) pairs* of the branch. Abhyankar [1] has shown the relationship between the characteristic pairs determined by expanding y in terms of x and those determined by expanding x in terms of y .

More complete discussions of the Newton polygon can be found in Chrystal's classic *Textbook of Algebra* [7, ch. 30, sect. 18–24] and Walker's *Algebraic Curves* [21, ch. 4, sect. 3]. More recently, Abhyankar [3] offers a proof of "Newton's Theorem" and some applications to algebraic geometry, while Brieskorn and Knorrer [4] provide a proof of the theorem, many examples, and details about characteristic pairs and their applications.

Newton's method for resolving affected equations can be a useful ingredient in undergraduate research projects. For example, one of my students, Tate Hilgefert, used Newton polygons to find a root of the general quintic equation [10]. As Bring [5] showed in 1786, a Tschirnhaus transformation can be found that reduces any quintic to the form $y^5 + y + x = 0$, where x is a radical expression in the coefficients of the original quintic [20]. Applying Newton's second method (and assisted by *Mathematica*) Hilgefert found the series solution

$$y = x + x^5 - 5x^9 + 35x^{13} - 285x^{17} + \dots$$

due to Eisenstein [20]. Thus, although the roots of

		+2, 10000000
		<u>-0, 00544852</u>
		2, 09455148
$2 + p = y$	y^3	+8 + 12p + 6pp + p ³
	-2y	-4 - 2p
	-5	-5
	summa	-1 + 10p + 6pp + p ³
$+0, 1 + q = p$	+p ³	+0, 001 + 0, 03q + 0, 3qq + q ³
	6pp	+0, 06 + 1, 2 + 6,
	+10p	+1 + 10,
	-1	-1
	summa	0, 061 + 11, 23q + 6, 3qq + q ³
$-0, 0054 + r = q$	+q ³	-0, 0000001 + 0, 000r &c
	+6, 3qq	+0, 0001837 - 0, 068
	11, 23q	-0, 060642 + 11, 23
	+0, 061	+0, 061
	summa	0, 0005416 + 11, 162r
$-0, 00004852 + s = r$		

Table 1. Newton’s First Diagram from the *Epistola prior*

the general quintic cannot be expressed as a finite algebraic combination of the coefficients, if series are permitted, the solutions can be found.

Appendix I. Portion of the *Epistola prior*

Most worthy Sir,

Though the modesty of Mr Leibniz, in the extracts from his letter which you have lately sent me, pays great tribute to our countrymen for a certain theory of infinite series, about which there now begins to be some talk, yet I have no doubt that he has discovered not only a method for reducing any quantities whatever to such series, as he asserts, but also various shortened forms, perhaps like our own, if not even better. Since, however, he very much wants to know what has been discovered in this subject by the English, and since I myself fell upon this theory some years ago, I have sent you some of those things which occurred to me in order to satisfy his wishes, at any rate in part...

The extractions of affected roots, of equations with several literal terms, resemble in form their extractions in numbers, but the method of Vieta and

our fellow-countryman Oughtred is less suitable for this purpose. Therefore I have been led to devise another, an example of which the following diagrams display, where the right-hand column exhibits the results of substituting in the middle column the values of y, p, q, r , etc. shown in the left-hand column. The first diagram displays the solution of this numerical equation, $y^3 - 2y - 5 = 0$; and here at the top of the column the negative part of the root, subtracted from the positive part, gives the actual root 2, 09455148; and the second diagram displays the solution of this literal equation, $y^3 + axy + a^2y + x^3 - 2a^3 = 0$.

In the first diagram the first term of the value of p, q, r in the first column is found by dividing the first term of the sum given in the line next above by the coefficient of the second term of the same sum, [as 1 by 10, or 0, 061 by 11, 23, and by changing the sign of the quotient]; and the same term is found in almost the same way in the second diagram. Here to be sure the chief difficulty is in finding the first term of the root; a general method by which this is effected I pass over here for the sake of brevity, as also some other things which tidy up the operation. And as there is not time here to explain the ways of abbreviating the process I shall merely say generally that the root of any equation once extracted can be

		$a - \frac{x}{4} + \frac{xx}{64a} + \frac{131x^3}{512aa} + \frac{509x^4}{16384a^3} \&c$
$a + p = y$	y^3	$a^3 + 3aap + 3app + p^3$
	$+axy$	$+aax + axp$
	$+aay$	$+a^3 + aap$
	$-x^3$	$-x^3$
	$-2a^3$	$-2a^3$
$-\frac{1}{4}x + q = p$	p^3	$-\frac{1}{64}x^3 + \frac{3}{16}xxq \&c$
	$+3app$	$+\frac{3}{16}aax - \frac{3}{2}axq + 3aqq$
	$+axp$	$-\frac{1}{4}aax + axq$
	$+4aap$	$-aax + 4aaq$
	$+aax$	$+aax$
	$-x^3$	$-x^3$
$+\frac{xx}{64a} + r = q$	$3aqq$	$+\frac{3x^4}{4096a} \&c$
	$+\frac{3}{16}xxq$	$+\frac{3x^4}{1024a} \&c$
	$-\frac{1}{2}axq$	$-\frac{1}{128}x^3 - \frac{1}{2}axr$
	$+4aaq$	$+\frac{1}{16}aax + 4aar$
	$-\frac{65}{64}x^3$	$-\frac{65}{64}x^3$
	$-\frac{1}{16}aax$	$-\frac{1}{16}aax$
		$+4aa - \frac{1}{2}ax) + \frac{131}{128}x^3 - \frac{15x^4}{4096a} \left(+\frac{131x^3}{512aa} + \frac{509x^4}{16384a^3} \cdot \right.$

Table 2. Newton's Second Diagram from the *Epistola prior*

kept as a rule for solving similar equations; and that from several such rules it is usually possible to form a more general rule; and that all roots, whether they be simple or affected, can be extracted in limitless ways, and on that account the simpler of the ways must always be considered.

Appendix II. Portion of the *Epistola posterior*

Most worthy Sir,

I can hardly tell with what pleasure I have read the letters of those very distinguished men Leibniz and Tschirnhaus. Leibniz's method for obtaining convergent series is certainly very elegant, and it would have sufficiently revealed the genius of its author, even if he had written nothing else. But what he has scattered elsewhere throughout his letter is most worthy of his reputation — it leads us also to hope for very great things from him. The variety of ways by which the same goal is approached has given me the greater pleasure, because three methods of arriving at series of that kind had already become known

to me, so that I could scarcely expect a new one to be communicated to us. One of mine I have described before; I now add another, namely, that by which I first chanced on these series—for I chanced on them before I knew the divisions and extractions of roots which I now use. And an explanation of this will serve to lay bare, what Leibniz desires from me, the basis of the theorem set forth near the beginning of the former letter. . .

What the celebrated Leibniz wants me to explain I have partly described above. But as to finding the terms p, q, r , in the extraction of an affected root, first I get p thus. Having described the right angle BAC , I divide its sides BA, AC into equal parts, and then draw normals dividing the angular space into equal parallelograms or squares, which I suppose to be designated by the dimensions of two indefinite kinds, say x and y , ascending in order from the end A , as is seen inscribed in Figure 1; where y denotes the root to be extracted and x the other indefinite quantity, from powers of which a series is to be constructed. Then, when some equation is proposed, I distinguish the parallelograms corresponding to each of its terms with some mark, and apply a

ruler to two or perhaps more of the marked parallelograms, of which one is the lowest in the left-hand column next AB , and others situated to the right of the ruler, while all the rest not touching the ruler lie above it. I pick out the terms of the equation distinguished by the parallelograms in contact with the ruler, and thence get the quantity to be added to the quotient.

B	x^4	x^4y	x^4y^2	x^4y^3	x^4y^4
	x^3	x^3y	x^3y^2	x^3y^3	x^3y^4
	x^2	x^2y	x^2y^2	x^2y^3	x^2y^4
	x	xy	xy^2	xy^3	xy^4
A	0	y	y^2	y^3	y^4

Figure 1.

Thus to extract the root y from

$$y^6 - 5xy^5 + (x^3/a)y^4 - 7a^2x^2y^2 + 6a^3x^3 + b^2x^4 = 0;$$

the parallelograms answering to the terms of this equation I denote by some mark * as in Figure 2. Then I apply the ruler DE to the lower of the places marked in the left-hand column, and rotate it from the lower to the higher to the right till it begins to reach likewise another or perhaps several of the remaining marked places. And I see that the places x^3, x^2y^2 and y^6 are thus reached. And so from the terms $y^6 - 7a^2x^2y^2 + 6a^3x^3$ as though equal to zero (and further reduced if desired to $v^6 - 7v^2 + 6 = 0$

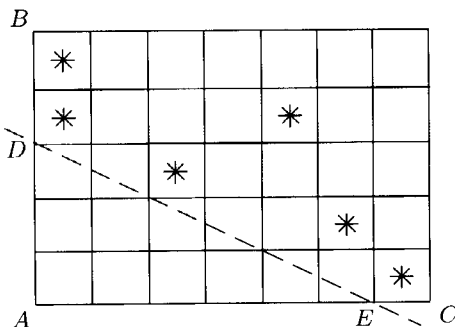


Figure 2.

by putting $y = v\sqrt{(ax)}$, I seek the value of y , and find four, namely

$$+\sqrt{(ax)}, -\sqrt{(ax)}, +\sqrt{(2ax)}, -\sqrt{(2ax)},$$

of which any one may be taken as the first term of the quotient, according as it has been decided to extract one or other of the roots.

Thus the equation

$$y^3 + axy + a^2y - x^3 - 2a^3 = 0,$$

which I solved in my former letter, gives

$$2a^3 + a^2y + y^3 = 0,$$

and hence $y = a$ very nearly. And so since a is the first term of the value of y , I put p for all the rest to infinity, and substitute $a + p$ for y . Here some difficulties will sometimes arise, but Leibniz I think will need no help to extricate himself from them. But the ensuing terms q, r, s are obtained, from the second and third equations and the rest, in the same way as the first term p from the first equation, only with less trouble, because the remaining terms of the value of y commonly result from dividing the term involving the lowest power of the variable x by the coefficient of the root p, q, r or $s. \dots$

References

1. S. S. Abhyankar, Inversion and invariance of characteristic pairs, *American Journal of Mathematics* 25 (1966) 77-86.
2. —, Historical ramblings in algebraic geometry and related algebra, *American Mathematical Monthly* 83 (1976) 409-48.
3. —, *Algebraic Geometry for Scientists and Engineers*, American Mathematical Society, Providence, 1990; lectures 12, 13, 21.
4. Egbert Brieskorn and Horst Knörrer, *Plane Algebraic Curves*, Birkhäuser, Boston, 1986; chapter 3, sections 8.3, 8.4.
5. E. S. Bring, *Meletemata quaedam Mathematica circa transformationem Aequationum Algebraicarum*, Lund University, Promotionschrift, 1786.
6. Florian Cajori, *A History of Mathematics*, Macmillan, New York, 1919.
7. G. Chrystal, *Textbook of Algebra*, vol. 2, Chelsea, New York, 1964. (Preprint of A. and C. Black, Edinburgh, 1900.)
8. William Dunham, Newton's (original) method — or — Though this be method, yet there is madness in't, address to the Mathematical Association of America, San Francisco, January, 1995.

9. G. Halphen, Étude sur les points singuliers des courbes algébriques planes, appendix to G. Salmon, *Higher Plane Curves* (French ed.), Dublin, 1852.
10. Tate Hilgefert, A power series solution to the general quintic, address to the Miami University Pi Mu Epsilon Conference, September, 1995.
11. Nick Kollerstrom, Thomas Simpson and 'Newton's method of approximation': An enduring myth, *British Journal for the History of Science* 25 (1992) 347–354.
12. Isaac Newton, De analysi, in D. T. Whiteside (ed.), *The Mathematical Papers of Isaac Newton*, Cambridge University Press, Cambridge, vol. 2, 1967–1981, 206–247.
13. —, De methodis serierum et fluxionum, in D. T. Whiteside (ed.), *The Mathematical Papers of Isaac Newton*, Cambridge University Press, Cambridge, vol. 3, 1967–1981, 32–353; pages 43–71.
14. —, Letter to Oldenburg, 13 June 1676, in H. W. Turnbull (ed.), *The Correspondence of Isaac Newton*, vol. 2, Cambridge University Press, Cambridge, 1960.
15. —, Letter to Oldenburg, 24 October 1676, in H. W. Turnbull (ed.), *The Correspondence of Isaac Newton*, vol. 2, Cambridge University Press, Cambridge, 1960.
16. V. A. Puiseux, Recherches sur les fonctions algébriques, *Journal of Mathematics Pure and Applied* 15 (1850) 365–480.
17. J. Raphson, *Analysis aequationum universalis*, London, 1690.
18. T. Simpson, *Essays . . . on mathematics*, London, 1740.
19. H. J. S. Smith, On the higher singularities of plane curves, *Proceedings of the London Mathematical Society* 6 (1873) 153–182.
20. John Stillwell, Eisenstein's footnote, *Mathematical Intelligencer* 17 (1995) 58–62.
21. Robert J. Walker, *Algebraic Curves*, Dover, New York, 1962; chapter 4, section 3. (Reprint of Princeton University Press, 1950.)
22. Richard S. Westfall, In defense of Newton: His biographer replies, *College Mathematics Journal* 25:3 (1994) 201–205.

A Contribution of Leibniz to the History of Complex Numbers

R. B. McCLENON

American Mathematical Monthly 30 (1923), 369–374

One of the most important and fascinating chapters in the history of mathematics is the development of the concept of complex numbers. Certain parts of this development have not yet been adequately treated by writers on the history of mathematics, and among these is to be mentioned the work of Leibniz.

It may be worth while to recall that neither the Hindu nor the Arabian algebraists, nor the medieval Europeans, had recognized any possibility of attaching a meaning to a square root of a negative number; indeed it was only the exceptional writer who recognized even *negative* roots of equations (for example, Leonardo of Pisa; see [5]). In the sixteenth century, Tartaglia and Cardan, in the formula for the roots of the cubic $x^3 + ax = b$, viz.,

$$x = \sqrt[3]{\frac{b}{2} + \sqrt{\frac{b^2}{4} + \frac{a^3}{27}}} + \sqrt[3]{\frac{b}{2} - \sqrt{\frac{b^2}{4} + \frac{a^3}{27}}}$$

noticed that in case $(b^2/4) + (a^3/27)$ were negative, the value of x would involve an “impossible” expression; and accordingly this case came to be known as the “irreducible case”, a term which persists down to the present time. Vieta (1540–1603), the greatest algebraist of his time, contented himself with working out a trigonometric solution for the cubic in this case (see [8], [1], [6]). Descartes, in connection with his “rule of signs”, mentioned the existence of imaginary roots in an algebraic equation, but did not enter upon any discussion of them [1].

It is now almost exactly 250 years since Leibniz, then a young man of 25, first entered upon the serious study of the possibility of getting some clear meaning out of these so-called “impossible” quantities. The inspiration for this work came to him

through the study of Bombelli’s *Algebra*, a standard work which had been published at Bologna in 1572 and reprinted in 1579. Leibniz was not at all satisfied with Bombelli’s discussion of the “Cardan” formula for the solution of the cubic equation, especially in the irreducible case. In a letter to Huygens [4], he expresses his dissatisfaction with Bombelli for not accepting Cardan’s formula as adequate in this case; and proceeds to make these three assertions: (1) that Cardan’s formula is universally valid, (2) that by means of this formula every cubic equation can be solved, and (3) that roots of all even degrees can be formed which contain imaginaries and yet which are themselves real. As an example of this last, Leibniz mentions that

$$\sqrt{1 + \sqrt{-3}} + \sqrt{1 - \sqrt{-3}} = \sqrt{6} \quad (1)$$

He also says in this same letter that he has found “a method for extracting, either exactly or approximately, the roots of binomials where imaginaries enter” [4]. In reply to this communication, Huygens expresses his astonishment at the relation (1) in these words: “The remark which you make concerning roots that can not be extracted, and containing imaginary quantities which when added together give none the less a real quantity, is surprising and entirely new. One would never have believed that

$$\sqrt{1 + \sqrt{-3}} + \sqrt{1 - \sqrt{-3}}$$

would make $\sqrt{6}$, and there is something hidden in this which is incomprehensible to us.” [4]

Leibniz evidently spent considerable time and effort on the question of the meaning of imaginary

expressions, and the possibility of securing reliable results by applying to them the ordinary laws of algebra; for Gerhardt found among Leibniz's papers a discussion of the solution of algebraic equations which, although undated, bears every evidence of having been written at about this time (1675). It is published in the *Briefwechsel*, pp. 550–564, and although it is one of the first significant documents in the history of complex numbers, it has not hitherto, so far as I know, been described by historians of mathematics. A rather full description of this paper will accordingly be worth while; and not alone because it has historical importance, but also because the clear-cut way in which Leibniz presents many of his points offers valuable suggestions to the teacher of the present day.

After stating the condition under which a quadratic equation will have real roots, Leibniz continues, "But if now a simple, that is, a linear equation, is multiplied by a quadratic, a cubic will result, which will have three real roots if the quadratic is possible, or two imaginary roots and only one real one if the quadratic is impossible." He then points out that it is exactly in the case where all three roots of the cubic are real that the difficulty in the use of Cardan's formula lies: the roots of

$$y^3 + qy - r = 0$$

being

$$y = \sqrt[3]{\frac{r}{2} + \sqrt{\frac{r^2}{4} + \frac{q^3}{27}}} + \sqrt[3]{\frac{r}{2} - \sqrt{\frac{r^2}{4} + \frac{q^3}{27}}}.$$

"How can it be", he says, "that a real quantity, a root of the proposed equation, is expressed by the intervention of an imaginary? For this is the remarkable thing, that, as calculation shows, such an imaginary quantity is only observed to enter those cubic equations that have no imaginary root, all their roots being real or possible, as has been shown by trisection of an angle, by Albert Girard and others [2], [6]. . . . This difficulty has been too much for all writers on algebra up to the present, and they have all said that in this case Cardan's rules fail."

Realizing clearly, then, the nature and difficulty of the problem, involving as its solution did a decisive step in advance of all his predecessors, Leibniz set to work to get to the bottom of the matter. He was led to the solution of the problem by an analogy in a similar situation. "It will be useful to mention how my mind was led to the solution of this problem. I

once came upon two equations of this kind:

$$x^2 + y^2 = b, \quad xy = c,$$

whence $x^2 = c^2/y^2$, and $c^2/y^2 + y^2 = b$ and $y^4 - by^2 + c^2 = 0$, or

$$y^2 = \frac{b}{2} + \sqrt{\frac{b^2}{4} - c^2},$$

$$y = \sqrt{\frac{b}{2} + \sqrt{\frac{b^2}{4} - c^2}}.$$

Substituting therefore this value of y^2 in $x^2 + y^2 = b$, I wrote

$$x^2 - \frac{b}{2} + \sqrt{\frac{b^2}{4} - c^2} = 0,$$

or

$$x = \sqrt{\frac{b}{2} - \sqrt{\frac{b^2}{4} - c^2}}.$$

But c was greater than b , and therefore

$$\sqrt{\frac{b^2}{4} - c^2}$$

was an imaginary quantity. However, I knew otherwise that the sum of the unknowns $x + y$ was a real quantity and equal to a certain line d , which puzzled me greatly, for inasmuch as I had deduced from the preceding calculation that

$$d = x + y = \sqrt{b/2 + \sqrt{b^2/4 - c^2}} + \sqrt{b/2 - \sqrt{b^2/4 - c^2}},$$

I did not understand how such a quantity could be real, when imaginary or impossible numbers were used to express it. I therefore began to retrace the steps of my calculation, suspecting an error; but in vain, for the result was always the same. At length it occurred to me to try this operation: put

$$d = \sqrt{b/2 + \sqrt{b^2/4 - c^2}} + \sqrt{b/2 - \sqrt{b^2/4 - c^2}} = A + B;$$

hence, squaring both sides,

$$d^2 = A^2 + B^2 + 2AB = b/2 + \sqrt{b^2/4 - c^2} + b/2 - \sqrt{b^2/4 - c^2} + 2c.$$

Therefore $d^2 = b + 2c$, and $d = \sqrt{b + 2c}$. Therefore, equating the two values of d ,

$$\sqrt{b + 2c} = \sqrt{b/2 + \sqrt{b^2/4 - c^2}} + \sqrt{b/2 - \sqrt{b^2/4 - c^2}}.$$

If we put $b = 2$ and also $c = 2$, there results

$$\sqrt{6} = \sqrt{1 + \sqrt{-3}} + \sqrt{1 - \sqrt{-3}}.$$

I do not remember to have noted a more singular and paradoxical fact in all analysis; for I think I am the first one to have reduced irrational roots, imaginary in form, to real values without extracting them.”

Thus Leibniz was led to what he called a *sixth arithmetical operation*, viz., the reduction of imaginary expressions to real form. He then proceeds to apply this operation to the Cardan form of the roots of a cubic equation. And first, he extends the principle of the preceding work with square roots to cube roots, as follows:

“Let $2b$ be a certain quantity: it can be written also in this way:

$$b + \sqrt{-ac} + b - \sqrt{-ac}.$$

For although $\sqrt{-ac}$ is an imaginary quantity, yet the sum is none the less real, since the imaginaries are destroyed. Let this formula be divided into two parts, the binomial $b + \sqrt{-ac}$ and the ‘apotome’¹ $b - \sqrt{-ac}$, and let us investigate the cube of each separately: the cube of $b + \sqrt{-ac}$ will be

$$b^3 - ac\sqrt{-ac} - 3bac + 3b^2\sqrt{-ac},$$

and the cube of $b - \sqrt{-ac}$ will be

$$b^3 + ac\sqrt{-ac} - 3bac - 3b^2\sqrt{-ac},$$

and therefore

$$2b = \sqrt[3]{b^3 - ac\sqrt{-ac} - 3bac + 3b^2\sqrt{-ac}} + \sqrt[3]{b^3 + ac\sqrt{-ac} - 3bac - 3b^2\sqrt{-ac}},$$

or

$$\sqrt[3]{b^3 - 3bac + \sqrt{-a^3c^3 + 6a^2c^2b^2 - 9b^4ac}} + \sqrt[3]{b^3 - 3bac - \sqrt{-a^3c^3 + 6a^2c^2b^2 - 9b^4ac}},$$

or

$$b + \sqrt{-ac} + b - \sqrt{-ac}.$$

“But if now from a binomial of this kind the cube root can always be extracted, as it can from this one, then certainly the imaginaries can always be removed from a binomial and an ‘apotome’ when

¹This is the designation used by Euclid in Book X of the *Elements*.

they are joined together. But since it can not always be extracted from a given expression in the form

$$\sqrt[3]{\frac{r}{2} + \sqrt{\frac{r^2}{4} + \frac{q^3}{27}}} + \sqrt[3]{\frac{r}{2} - \sqrt{\frac{r^2}{4} - \frac{q^3}{27}}},$$

such as cubic equations give, that is, since the given quantity $r/2$ can not always be separated into two, $b^3 - 3bac$, nor the given quantity $(r^2/4) - (q^3/27)$ into three, $-a^3c^3 + 6a^2c^2b^2 - 9b^4ac$, without another equation, equally as difficult as the given one, therefore it happens that we can not always eliminate the imaginaries from real quantities.

“But it will be useful to give examples in rational numbers. Take the equation, which also Albert Girard used [2]:

$$x^3 - 13x - 12 = 0,$$

whose true root is 4. From the formulas of Scipio Ferro or Cardan,

$$x = \sqrt[3]{6 + \sqrt{\frac{-1225}{27}}} + \sqrt[3]{6 - \sqrt{\frac{-1225}{27}}}.$$

I will prove that this expression is correct and real, and must be admitted. Put

$$x = 2 + \sqrt{-\frac{1}{3}} + 2 - \sqrt{-\frac{1}{3}},$$

and certainly x will be equal to 4, as the equation postulated. Now let us see if the Cardan formula can be derived from this. Certainly by cubing and applying the above formula

$$b + \sqrt{-ac} + b - \sqrt{-ac}$$

to this, making $b = 2$, and $ac = \frac{1}{3}$, we shall have for the cube of $2 + \sqrt{-\frac{1}{3}}$ this formula:²

$$+8 - 3 \cdot 2 \cdot \frac{1}{3} + \sqrt{-\frac{1}{27}} + 6 \cdot \frac{1}{9} \cdot 4 - 9 \cdot 16 \cdot \frac{1}{3}$$

or, adding up,

$$6 + \sqrt{\frac{-1225}{27}}.$$

In the same way the cube of $2 - \sqrt{-\frac{1}{3}}$ will be

$$6 - \sqrt{\frac{-1225}{27}}$$

²Leibniz actually uses commas to indicate multiplication; he later introduced the dot which has been universally adopted. See [7].

and hence

$$\sqrt[3]{6 + \sqrt{\frac{-1225}{27}}}$$

will be $2 + \sqrt{-\frac{1}{3}}$ and

$$\sqrt[3]{6 - \sqrt{\frac{-1225}{27}}}$$

will be $2 - \sqrt{-\frac{1}{3}}$ and, by joining the binomial to the ‘apotome’ x , or

$$\sqrt[3]{6 + \sqrt{\frac{-1225}{27}}} + \sqrt[3]{6 - \sqrt{\frac{-1225}{27}}}$$

will be the same as

$$2 + \sqrt{-\frac{1}{3}} + 2 - \sqrt{-\frac{1}{3}},$$

that is, will be 4, as was proposed to show.”

Leibniz adds an example where a negative number (-6) is a root of the cubic, $x^3 - 48x - 72 = 0$, and establishes the fact that

$$\sqrt[3]{36 + \sqrt{-2800}} + \sqrt[3]{36 - \sqrt{-2800}} = -6.$$

He finally takes the bull by the horns, substitutes in the cubic equation $x^3 - qx - r = 0$ the expression for x given by the Cardan formula, and shows by actually carrying out the algebraic reductions that the equation is thus satisfied.

The rest of the memoir is devoted to a discussion of the great difficulty of extending the methods of solution to the 5th, 6th, and higher degree equations with emphasis upon the necessity of doing this. The concluding sentences are as follows [4]: “For this evil I have found a remedy and obtained a method, by which without experimentation the roots of such

binomials can be extracted, imaginaries being no hindrance, and not only in the case of cubics but also in higher equations. This invention rests upon a certain peculiarity which I will explain later. Now I will add certain rules derived from the consideration of irrationals (although no mention is made of irrationals), by which a rational root can easily be extracted from them.”

Here the manuscript breaks off; no doubt Leibniz became convinced that he could not carry his “method” as far as he had at first supposed, and thus the essay was left unfinished. But the influence of this work of Leibniz is seen in the writings of Tschirnhausen on the one hand and of John Bernoulli on the other, each of whom received stimulation and valuable assistance from Leibniz in the field of algebra. Thus this particular memoir on complex numbers, although remaining unpublished for two centuries, is an interesting and important document in the history of mathematics.

References

1. Cantor, *Vorlesungen über Geschichte der Mathematik*, Leipzig, 1900.
2. Girard, *Invention nouvelle en l'Algèbre*, Amsterdam, 1629.
3. Heath, *History of Greek Mathematics*, Oxford, 1921.
4. Leibniz, *Der Briefwechsel von Gottfried Wilhelm Leibniz mit Mathematikern*, ed. C. J. Gerhardt, Berlin, 1899.
5. Leonardo of Pisa, *Scritti* (ed. Boncompagni), Rome, 1857.
6. Montucla, *Histoire des Mathématiques*, Paris, 1799.
7. Tropicke, *Geschichte der Elementarmathematik*, Berlin and Leipzig, 1921.
8. Zeuthen, *Geschichte der Mathematik im XVI und XVII Jahrhundert*, Leipzig, 1903.

Functions of a Curve: Leibniz's Original Notion of Functions and Its Meaning for the Parabola

DAVID DENNIS and JERE CONFREY

College Mathematics Journal 26 (1995), 124–131

When the notion of a function evolved in the mathematics of the late seventeenth century, the meaning of the term was quite different from our modern set theoretic definition, and also different from the algebraic notions of the nineteenth century. The main conceptual difference was that curves were thought of as having a primary existence apart from any analysis of their numeric or algebraic properties. Equations did not create curves, curves gave rise to equations. When Descartes published his *Geometry* [10] in 1637, he derived for the first time the algebraic equations of many curves, but never once did he create a curve by plotting points from an equation. Geometrical methods for drawing each curve were always given first, and then by analyzing the geometrical actions involved in the curve drawing apparatus he would arrive at an equation that related pairs of coordinates (not necessarily at right angles to each other) [20]. Descartes used equations to create a taxonomy of curves [17].

This tradition of seeing curves as the result of geometrical actions continued in the work of Roberval, Pascal, Newton, and Leibniz. Descartes used letters to represent various lengths but did not create any specific system of names. Leibniz, who introduced the term *function* into mathematics [2], considered six different functions associated with a curve, i.e., line segments or lengths that could be determined from each point on a curve relating it to a given line or axis. He gave them the names abscissa, ordinate, tangent, subtangent, normal, and subnormal. These six are shown in Figure 1 for the curve RP , relative to the axis AO . The line PO is perpendicular to AO . The line PT is tangent to the curve at P , and the line PN is perpendicular to PT .

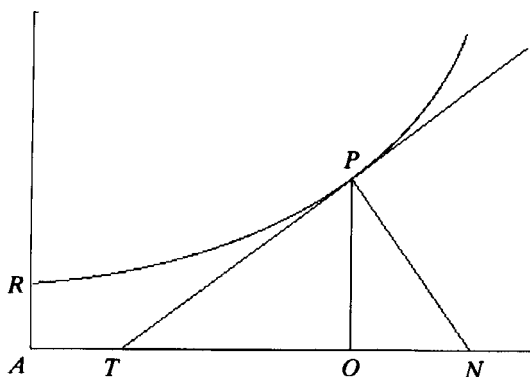


Figure 1. PO ordinate; AO abscissa; PT tangent; OT subtangent; PN normal; ON subnormal

It is important to note here that the curve and an axis must exist before these six functions can be defined. In this definition, the abscissa and ordinate may at first seem to be a parametric representation of the curve, but this is not the case. No parameter, such as time or arc length, is involved. The setting is entirely geometric. From the geometric point P , the line segments (functions) are defined relative to the axis AO . *Abscissa* is Latin for “that which is cut off,” i.e., a piece of the axis AO is cut off. By cutting off successive pieces of the axis, the curve gives us an ordered series of line segments PO as P moves along the curve. Hence the term *ordinate*.

It should also be noted here that all of these functions of a point P on a given curve are defined without reference to any particular unit of measurement. They are line segments. Leibniz, of course, like Descartes, wanted to introduce quantification and analyze the properties of curves algebraically,

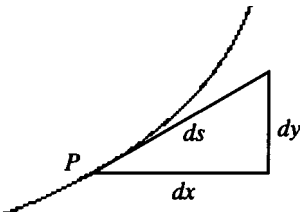


Figure 2.

but since the definition of the functions is geometric he could postpone the choice of a unit until an appropriate one could be found for the curve at hand. The advantage of this will emerge in our discussion of the parabola.

Since angles TPN , POT , and PON are right angles, the triangles TOP , RON , and TPN are all similar. This configuration will be familiar to geometers as the construction of a geometric mean between ON and OT , the mean being OP .

Inspired by the work of Pascal, Leibniz saw a fourth triangle which was similar to the three mentioned above [2], [5], [11]. This was the *infinitesimal* or *characteristic* triangle (see Figure 2), used by Pascal to integrate the sine function [21]. Leibniz viewed a geometric curve as made up of infinitely small line segments which each had a particular direction. He perceived the utility of this concept in Pascal's work and it became one of the primary notions in his development of a system of notation for calculus. Although many modern mathematicians avoid this conception, it is still used as an important conceptual device by engineers. Figure 2 still appears in calculus books because it conveys an important meaning, especially to those who use calculus for the analysis of physical or mechanical actions. (With the invention, early in this century, of the calculus of differentials as linear functions on the tangent lines to the curve, Leibniz's fundamental insight was made rigorous without recourse to "infinitesimals" [18].)

Leibniz saw great significance in the triangles of Figure 1 because they were large and visible yet similar to the unseen characteristic triangle. This finding of large triangles that are similar to infinitesimal ones is a theme that runs through many of the most important works of Leibniz [5], [8], [11]. From Figures 1 and 2, the similarity relations tell us that

$$\frac{dy}{dx} = \frac{PO}{OT} = \frac{ON}{PO}.$$

Let us look at how this system works in the case of the parabola. We must first have a way to draw

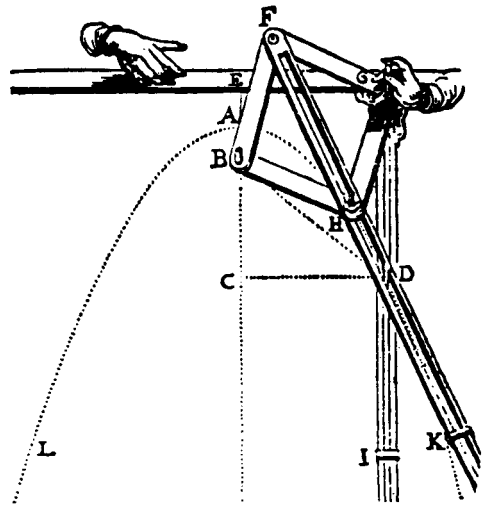


Figure 3.

a parabola. Everything begins with the existence of a curve. Figure 3 shows a linkage that will draw parabolic curves. This figure comes from the work of Franz Van Schooten (1615–1660) [23, p. 359], whose extensive commentaries on Descartes' *Geometry* were widely read in the seventeenth century [22]. Because his works supplied many of the details Descartes omitted they were in fact more popular than the *Geometry* itself.

This apparatus constructs the parabola from the familiar focus/directrix definition. That is, the parabola is the set of points equidistant from a point and a line. The ruler GE is the directrix and the point B is the focus. Four equal-length links create a movable rhombus $BFGH$ which guarantees that FH will always be the perpendicular bisector of BG as G moves along the ruler. GI is a movable ruler that is always perpendicular to the directrix EG . The point D is the intersection of FH and GI as the point G moves along the directrix. Hence at all positions $BD = GD$, and hence D traces a parabola with focus B and directrix EG .

This construction can be simulated on a computer using the software *Geometer's Sketchpad* [14]. This software allows one to define a perpendicular bisector so the rhombus is unnecessary. One can either drag a point along the directrix or have the computer animate such a motion. Figure 4 was made using this software. The point F is the focus, and the point S is moving along the directrix. BP is the perpendicular bisector of FS , SP is always perpendicular to the directrix, and the intersection point P traces a parabola.

invariant of this curve-drawing action and can be seen as the natural choice of a unit for this curve. As it turns out, the subnormal ON is always equal to the distance between the focus and the directrix of the parabola. Thus it is a natural unit. Using the subnormal as a unit, the equation of the parabola becomes $x = y^2/2$, i.e., the common integral form of the parabola as the accumulated area under the line $x = y$. It is in this form that the parabola most often appears in the table interpolations of John Wallis and Isaac Newton [9].

One way to prove that the subnormal is constant is to show that it always equals the distance between the focus and the directrix. Looking at Figure 5, we see that SF and PN are both perpendicular to BP , so triangles SCF and PON are congruent; hence $ON = CF$.

In order to prove that the vertex A is always the midpoint of the subtangent OT , one can establish that triangles TBA and PBK are congruent. They are clearly similar, but since B is the midpoint of SF it is also the midpoint of AK , so they are congruent. Hence $TA = KP = AO$.

Lastly, one might ask: How can we be sure that the line BP is always tangent to the parabola? That is to say, how can we be sure that each instance of the line BP intersects the parabola in only one point? Let $Q \neq P$ be a point on BP , and let R be the foot of the perpendicular from Q to the directrix CS . Since R is the closest point to Q on the directrix, $QR < QS$. Since BP is the perpendicular bisector of SF , $QS = QF$. Hence $QR < QF$ and Q cannot be on the parabola, being closer to the directrix than to the focus. One could also check the tangency of BP analytically by writing the equation of the parabola and the line BP using the same coordinate system and then solving the two equations simultaneously, arriving at a quadratic equation with one repeated root. This is the method that Descartes developed for finding tangents; i.e., tangency occurs when repeated roots appear in the simultaneous solutions.

These two invariant properties of the parabola were never mentioned (so far as we know) in the published work of Leibniz. The fact that the vertex is the midpoint of the subtangent was demonstrated by Apollonius [1]. The fact that the subnormal is constant is credited to L. Euler, who expanded and popularized the ideas of Leibniz [7]. They both appear in Book 2 of Euler's most famous textbook, the *Introduction to Analysis of the Infinite* [12]. This book, published in 1748, was the first modern precalculus textbook and, along with its sequels on dif-

ferential and integral calculus, did much to standardize curriculum and notation. Nearly all of the topics in our modern precalculus books are contained in Euler's book, but what is missing from our modern treatments is the bold empirical spirit of Euler's investigations, as well as most of his more advanced geometry and infinite series. Euler says in the preface to his text that he presents many questions that can be more quickly resolved using calculus. He insists, however, that when students rush into calculus too rapidly they become confused, because they lack the experiential basis (both geometric and algebraic) upon which calculus is built.

The parabola example demonstrates how much can be found using only basic geometry combined with empirical investigation. By letting the configuration move, we create a situation where algebra evolves naturally from geometry. Too often in our schools we find our geometry curriculum static and isolated from other topics, especially algebra. Two-column geometry proofs provide a shadow of Euclid, but they cannot provide the dynamic experience that leads to an understanding of functions and calculus. An important philosophical prerequisite for understanding calculus is the belief that geometry and algebra are consistent with each other, and historically this belief did not come easily [4]. This belief is too often tacitly assumed in our classrooms. In order for students to comprehend and appreciate this they must first be allowed to experience doubt as to whether a geometric result will be confirmed by an arithmetic result [8]. With modern software, computers can now readily simulate moving geometry, and this experience can be very compelling. For some, an empirical experience based on mechanical devices or paper folding can be even more compelling.

For the reader who wishes to attempt this kind of analysis on other curves, we offer the following tantalizing tidbits. If the directrix in the above construction is a circle instead of a line, then one can draw both hyperbolas and ellipses with their tangents [8], [23]. Paper folding also works [13], [19]. In the case of the hyperbola, if a tangent line at a point P is extended until it intersects the asymptotes at points A and B , then P will always be the midpoint of the segment AB . This little-known theorem is in Euler [12] but goes back to Apollonius [1]. As an empirical observation this can lead in many analytic directions. For example, the derivative of $y = 1/x$ can immediately be seen to be $-1/x^2$. Check it out! (Similar methods can be applied to draw planetary orbits; see the wonderful article by A. Lenard [16].)

Exercise. We have shown that parabolas have constant subnormals. What curves have constant subtangents? (Answer precedes reference list.)

In order to have the kind of empirical experience that Lakatos [15] suggests is fundamental to mathematical discovery, people should be encouraged to design, build, and explore their own devices and computer simulations. Some experience with mechanical devices can greatly aid many students as they attempt to master the use of software like *Geometer's Sketchpad*. All algebraic curves, for example, can be drawn with linkages [3]; some are easily built and others are best simulated. The border between mathematics, simulation, and mechanical engineering can become quite fuzzy. In such a setting geometry and algebra complement, validate, and empower one another without forming a hierarchy.

After many years of working in mathematics education at all levels, we have come to believe that effective educational practice must involve people in a balanced dialogue between “grounded activity” and “systematic inquiry” [6]. This discussion of the parabola provides an excellent example of such a dialogue.

Answer to Exercise. Exponential curves $y = y_0 e^{kt}$. For a discussion of this question and many others like it, see [8].

References

1. Apollonius of Perga, *Treatise on Conic Sections* (vol. II of *Great Books of the Western World*), Encyclopedia Britannica, Chicago, 1952.
2. V. I. Arnol'd, *Huygens & Barrow, Newton & Hooke*, Birkhäuser Verlag, Boston, 1990.
3. I. I. Artobolevskii, *Mechanisms for the Generation of Plane Curves*, Macmillan, New York, 1964.
4. F. Cajori, Controversies on mathematics between Wallis, Hobbes, and Barrow, *Mathematics Teacher* 22:3 (1929) 146–151.
5. J. M. Child, *The Early Mathematical Manuscripts of Leibniz*, Open Court, Chicago, 1920.
6. J. Confrey, The role of technology in reconceptualizing functions and algebra, in Joanne Rossi Becker and Barbara J. Pence, eds., *Proceedings of the Fifteenth Annual Meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education*, Pacific Grove, CA, October 17–20, vol. I, The Center for Mathematics and Computer Science Education at San Jose State University, San Jose, CA, 1993, 47–74.
7. J. L. Coolidge, *A History of Conic Sections and Quadratic Surfaces*, Dover, New York, 1968.
8. D. Dennis, *Historical perspectives for the reform of mathematics curriculum: Geometric curve drawing devices and their role in the transition to an algebraic description of functions*, Unpublished doctoral dissertation, Cornell University, 1994.
9. D. Dennis and J. Confrey, The creation of binomial series: A study of the methods and epistemology of Wallis, Newton, and Euler, presented at the AMS-CMS-MAA joint meetings, Vancouver, BC, August 15–19, 1993. Available from authors.
10. R. Descartes, *The Geometry*, Open Court, LaSalle, IL, 1952.
11. C. H. Edwards, *The Historical Development of Calculus*, Springer-Verlag, New York, 1979.
12. L. Euler, *Introduction to Analysis of the Infinite* (2 vols.), Springer-Verlag, New York, 1988, 1990.
13. M. Gardner, *Penrose Tiles to Trapdoor Ciphers*, Freeman, New York, 1989.
14. *Geometer's Sketchpad*TM (version 2.1), N. Jackiw, Key Curriculum Press, Berkeley, CA, 1994.
15. I. Lakatos, *Proofs and Refutations: The Logic of Mathematical Discovery*, Cambridge University Press, New York, 1976.
16. A. Lenard, Kepler orbits, more geometrico, *College Mathematics Journal* 25:2 (1994) 90–98.
17. T. Lenoir, Descartes and the geometrization of thought: The methodological background of Descartes' geometry, *Historia Mathematica* 6 (1979) 355–379.
18. M. E. Munroe, *Calculus*, Saunders, Philadelphia, 1979, p. 92.
19. T. S. Row, *Geometric Exercises in Paper Folding*, Dover, New York, 1966.
20. E. Smith, D. Dennis, and J. Confrey, Rethinking functions, Cartesian constructions, in S. Hills, ed., *The History and Philosophy of Science in Science Education, Proceedings of the Second International Conference on the History and Philosophy of Science and Science Education*, vol. 2, The Mathematics, Science, Technology and Teacher Education Group, Queens University, Kingston, Ontario, 1992, 449–466.
21. D. J. Struik, *A Source Book in Mathematics, 1200–1800*, Harvard University Press, Cambridge, MA, 1969.
22. J. van Maanen, Seventeenth century instruments for drawing conic sections, *Mathematical Gazette* 76: 476 (1992) 222–230.
23. F. Van Schooten, *Exercitationum Mathematicorum libri quinque*, Leiden, 1657 (original edition in the rare books collection of Cornell University, Ithaca, New York).

Afterword

Further information on the development of the calculus can be found in several good books. Margaret Baron's *The Origins of the Infinitesimal Calculus* [2] deals with many of the methods of the calculus up to the time of Newton and Leibniz. C. H. Edwards' *The Historical Development of the Calculus* [7] also shows how mathematicians calculated solutions to problems, but covers in more detail the work of Newton, Leibniz, and their successors. The classic work by Carl Boyer, *The History of the Calculus and its Conceptual Development* [4], concentrates more on the central ideas of the calculus rather than the technical details.

The mathematical work of Newton is available in English translation in the magnificent set, *The Mathematical Papers of Isaac Newton* [14], edited by D. T. Whiteside. In addition, there is a new English translation and commentary on Newton's *Principia* [10], by I. Bernard Cohen and Anne Whitman. Among the many other books which help the reader understand Newton's masterwork are Niccolò Guicciardini's *Reading the Principia* [9] and Dana Densmore's *Newton's Principia: The Central Argument* [6]. Both of these books deal further with the question that Pourciau considers, along with much other material. Leibniz's works are unfortunately not all available in English, but some of his early manuscripts have been collected and translated by J. M. Child in *The Early Mathematical Manuscripts of Leibniz* [5]. For an introduction to either man's work, it might be best to look through one of the standard biographies: *Never at Rest* [13] by Richard Westfall for Newton, and *Leibniz: A Biography* [1] by Eric Aiton for Leibniz.

There are a number of more specialized works on the development of the ideas of the calculus discussed in these articles. Judith Grabiner's *The Origins of Cauchy's Rigorous Calculus* [8] expands on the ideas in her paper. Roberval's work can be seen in detail in *A Study of the Traité des Indivisibles of Gilles Persone de Roberval* [12] by Evelyn Walker. Carl Boyer's *History of Analytic Geometry* [3] gives lots of detail of various aspects of this history. More information on Gregory can be found in the *James Gregory Tercentenary Memorial Volume* [11], edited by H. W. Turnbull.

References

1. Eric Aiton, *Leibniz, A Biography*, Adam Hilger Ltd., Bristol, 1985.
2. Margaret Baron, *The Origins of the Infinitesimal Calculus*, Pergamon Press, Oxford, 1969.
3. Carl Boyer, *History of Analytic Geometry*, Scripta Mathematica, New York, 1956.
4. Carl Boyer, *The History of the Calculus and its Conceptual Development*, Dover, New York, 1959.
5. J. M. Child, *The Early Mathematical Manuscripts of Leibniz*, Open Court, Chicago, 1920.
6. Dana Densmore, *Newton's Principia: The Central Argument*, Green Lion Press, Santa Fe, 1995.
7. C. H. Edwards, *The Historical Development of the Calculus*, Springer, New York, 1979.
8. Judith Grabiner, *The Origins of Cauchy's Rigorous Calculus*, MIT Press, Cambridge, 1981.

9. Niccolò Guicciardini, *Reading the Principia: The Debate on Newton's Mathematical Methods for Natural Philosophy from 1687 to 1736*, Cambridge University Press, 1999.
10. Isaac Newton, *The Principia*, newly translated by I. Bernard Cohen and Anne Whitman. University of California Press, Berkeley, 1999.
11. H. W. Turnbull (ed.), *James Gregory Tercentenary Memorial Volume*, G. Bell and Sons, London, 1939.
12. Evelyn Walker, *A Study of the Traité des Indivisibles of Gilles Persone de Roberval*, Columbia University Press, New York, 1932.
13. Richard Westfall, *Never at Rest*, Cambridge University Press, 1980.
14. D. T. Whiteside, *The Mathematical Papers of Isaac Newton*, Cambridge University Press, 1967–1981.



**The
Eighteenth
Century**

Foreword

Newton and Leibniz invented calculus in the late seventeenth century. The following century saw its continued development, so many of the articles in this section deal with aspects of the calculus. But since the towering figure in the eighteenth century is Leonhard Euler, much of this section deals with aspects of his work as well, both in analysis and in number theory.

The opening article of this section, however, deals with neither of these subjects. Although Brook Taylor is best known for his 1715 work *Methodus Incrementorum Directa et Inversa*, in which he discusses the Taylor series expansion of a function, in that same year he also published a book entitled *Linear Perspective*, in which he teaches methods for artists to represent three-dimensional objects on two-dimensional canvases. This work went through several editions and was translated into French and Italian, but in general proved too abstract for the artists to whom it was addressed. P. S. Jones analyzes several interesting ideas in the book, especially those that appeared in the first edition but were removed in later editions.

Evidently, one of the reasons Taylor's work was not as well received on the Continent as it might have been was that it was caught in the dispute between the followers of Newton and those of Leibniz on the origins of the calculus. Another important British author who was caught in that controversy was Colin Maclaurin. His massive *Treatise on Fluxions*, which aimed to justify and extend Newton's version of the calculus, is generally thought to have had little influence on the Continental developers of analysis. But as Judith Grabiner shows in the next article, this idea is entirely mistaken; Maclaurin's work was read, understood, and used by such mathematicians as Euler and Lagrange. In fact, many of Maclaurin's ideas in this work had direct influence on the subsequent work of d'Alembert on limits, Euler and Lagrange on series, and Clairaut on the gravitational attraction of ellipsoids, among much else.

One of the reasons that Maclaurin wrote his *Treatise* was to answer Bishop Berkeley's criticisms of Newton's approach to the calculus. In the next article, Florian Cajori discusses Maclaurin's answers to Berkeley and the responses of James Jurin, Benjamin Robins, and several others, concluding with the work of Robert Woodhouse in 1805. Woodhouse was the immediate predecessor of the group of young mathematicians at Cambridge who formed the Analytical Society, which aimed to bring the Continental approach to calculus to Britain to replace the increasingly sterile Newtonian version.

Among the most important developers of Leibniz's calculus on the European continent were Jakob and Johann Bernoulli. In the next article, William Dunham deals with Jakob's proof of the divergence of the harmonic series, a proof that Jakob attributes to his younger brother and which is quite different from the standard proof dating back to Oresme in the fourteenth century. Dunham mentions that Jakob attempted afterwards to sum the series of the reciprocals of the squares of the

integers, without success. This sum was finally found by Leonhard Euler, to whom the remainder of the articles in this section are devoted.

J. J. Burckhardt was one of the editors of the Euler memorial volume, published in Switzerland in 1983 to commemorate the two-hundredth anniversary of Euler's death. Here he summarizes some of Euler's accomplishments. In particular, he deals with some of Euler's lesser-known contributions, in the fields of physics and astronomy. These include the text, the *Dioptrica*, on the principles of optics and their application to the construction of optical instruments, and his work on perturbation theory, especially as applied to the movements of the planets.

Among Euler's numerous achievements was his analysis of the relationship between logarithms and exponentials in his 1748 *Introduction to Analysis of the Infinite*. In the next article, Julian Lowell Coolidge traces the history of this relationship before Euler, beginning with the Greek work on the rectangular hyperbola, including material from Gregory of St. Vincent, Christiaan Huygens, John Wallis, Isaac Newton, and Gottfried Leibniz. In particular, he notes that until the work of Euler, most authors did not consider logarithms as exponents, so there was little consideration of the base of a logarithmic system. It was Euler who gave us the modern definition and succeeded in calculating e to numerous decimal places.

Another great achievement of Euler was his idea that all kinds of functions needed to be admitted into analysis. Although initially Euler considered only functions that were 'analytic expressions', his work on the problem of the vibrating string led him to reconsider. Jesper Lützen explains Euler's vision of generalizing analysis to more general functions, especially to functions of two or more variables, and how his vision was realized in the work of L. Schwartz on distributions in the twentieth century.

The final three articles in this section deal with situations in which Euler was not completely successful in solving problems he had set for himself. It is important for our students to see that even the great Euler could fail, but in his failures there were always the seeds for the successful solution of his problem by later mathematicians. William Dunham addresses one of Euler's 'failures', his unsuccessful attempt to prove the fundamental theorem of algebra. Yet, as Dunham shows, an examination of the details of Euler's attempt allows even high-school students to grasp the meaning of the theorem along with Euler's perfectly correct treatment of the fourth- and fifth-degree cases.

Anthony Ferzola then looks at Euler's definition and use of differentials. What are these mysterious dx 's? Euler thinks of them as 'zeros', but with the property that the ratio of any two of them needs to be determined. Using this basic idea, we then follow Euler's determination of the differentials of products and quotients, of logarithms, and of trigonometric functions. Ferzola also discusses Euler's attempt to justify the change of variable formula in double integration by use of differentials. It was not until Elie Cartan applied Grassmann's exterior product to the algebra of differential forms that this formula was justified in the way that Euler desired.

Finally, Harold Edwards leads us through Euler's detailed computations with integers which ultimately led to his statement, but not proof, of a result equivalent to the famous quadratic reciprocity theorem. In this article, as in the previous one, we see Euler at work, doing lots of computations, making conjectures, correcting errors, and finally coming up with accurate conclusions. But although Euler did not succeed in proving his conjectures, his challenge was ultimately taken up by Gauss, who was so enamored by the quadratic reciprocity theorem that he published six proofs of it.

Brook Taylor and the Mathematical Theory of Linear Perspective

P. S. JONES

American Mathematical Monthly 58 (1951), 597–606

One can distinguish four overlapping and interrelated periods in the development of the mathematical theory of linear perspective:

- (1) the “prehistory” period in which, for example, the Greeks are reported to have made some use of perspective drawing in their theater,
- (2) the 15th and 16th century period of the origin of the theory with the artists-architects-engineers of the Renaissance (Brunelleschi, Franceschi, Alberti, and da Vinci),
- (3) a period of geometrical expositions typified by the works of del Monte and Stevin in the 17th century, and, finally,
- (4) the period of a generalized, complete, and even abstract theory.

This last period falls largely in the 18th century and is typified by the work of William Jacob Gravesand in Holland, Humphrey Ditton and Brook Taylor in England, and of the Alsatian (he was born in Mülhausen in the period when it was allied with Switzerland) mathematician Johann Heinrich Lambert.

Of these, the work of Brook Taylor was certainly the most widely translated and reproduced, although the later work of Lambert rivals it in interest and perhaps in its total effect [1].

Brook Taylor published only two books in his lifetime of 46 years. Both of these appeared in 1715 when he was 30, and both of them exerted wide influence. He is, of course, best known for his *Methodus Incrementorum Directa et Inversa* in which appears the well known expansion of $f(x + h)$ which bears his name.

The other book was

*LINEAR PERSPECTIVE OR, A
New METHOD
Of Representing justly all manner of
OBJECTS as they appear to the EYE
IN ALL
SITUATIONS.
A Work necessary for PAINTERS,
ARCHITECTS, & c. to Judge of, and
Regulate Designs by.*

This work is today only rarely and sparingly referred to in histories of either mathematics or art. This alone is of interest in view of a study which shows that the original appeared in four editions (or five, if Ware’s revision be counted), the latest as recent as 1811, that it appeared in three translations, one French and two Italian, and that Taylor’s English disciples in perspective number nine and were responsible for twelve books and twenty-two editions from 1715 through 1888 [2]. By disciples I here mean men who used Taylor’s name in the titles or body of their own works which works in turn followed more or less closely Taylor’s sequence and method.

One reason for this lack of recognition of Taylor’s *Perspective* is perhaps the same defect as that upon which John Bernoulli is said to have seized when, according to Taylor’s grandson and biographer, he called the book “abstruse to all and unintelligible to artists for whom it was more especially written” [3]. I have not found these exact words but it is quite likely both that Bernoulli said them and that one must discount them a little because of the heated and sharp nature of the controversy carried on by these

two men in the pages of *Acta Eruditorum* and the *Philosophical Transactions*, beginning with a letter by Bernoulli in 1716 [4]. This controversy was over priority and the proper recognition of sources used in both Taylor's *Methodus Incrementorum* and also in his publications on the vibrating string and on an isoperimetric problem. More than this, however, it was a part of the continuation by their partisans of the Newton-Leibniz Controversy which was not always conducted in a fair and rational vein.

Taylor himself, however, recognized the excessive conciseness and abstractness of his first book on perspective when he expanded it from 42 pages to the 70 found in the second, or 1719 edition, and when he added a few plates showing the application of his method to actual drawings of physical objects in addition to the purely geometric diagrams of the first edition.

A later evaluation by Monge and Lacroix is interesting. In An 9 (1801) they recommended to the Académie des Sciences of the Institut de France that it not sanction the publication of a French translation of the first edition by B. Lavite [5]. In the introduction to their report, however, they remarked that Taylor's work was "distinguished from a crowd of others dealing with perspective by its originality and the fruitfulness of the principle upon which it was based." They also termed it "elegant", "expeditious", and "not lacking in a sort of generality". They explained that they did not favor printing it in spite of this for two reasons; namely, that additional work or study of perspective was unnecessary for those who already knew "Stéréotomie", and that Taylor's work was too geometrical for most artists who were not versed in Stéréotomie. This seems a fair and rational evaluation when one recalls that it was made by the founder of descriptive geometry and one of his followers.

More recently Julian Coolidge has referred to Taylor's work as the "capstone of the whole edifice" of perspective [6]. In spite of this and the fact that Gino Loria also has paid some attention to Taylor's work [7], the tabulation of editions, translations and extensions which is noted above and detailed in the notes has not been made before, nor is there a discussion of the first or 1715 edition available since later writers on perspective used the second edition and the historians have used either it or versions still more remote from the original edition.

In this paper are presented only three of the items of especial interest which appeared in the 1715 edition but not in later editions. First, however, it will

be helpful to note that Taylor found it necessary to, as he said, "Consider this subject entirely anew." To this end he gave new terms, four axioms (in the 1719 edition), and then developed his theory in a formal and rigorous fashion with theorems, corollaries, problems, and proofs. He defined the "vanishing line" of any "original plane" to be the intersection with the picture plane of a plane through the eye of the beholder parallel to the original plane. This means that his basic three dimensional diagram as shown in Figure 1 (Plate 1 of Taylor's book) consisted of four planes parallel in pairs, the picture plane, the "directing plane" through the eye of the beholder and parallel to the picture, the original plane, and the plane through the eye parallel to it. The "vanishing point" of any "original line" is the intersection with the picture of a line through the eye parallel to the original line. Since the intersection of any original line with the picture is its own perspective, it follows as "PROPOSITION I, THEOREM 1" that "*The representation of a Line is Part of a Line passing thro' the intersection and Vanishing Point of the Original Line.*"

The above discussion of Taylor's terminology and Theorem 1 indicates three things about his work; namely, his formal, mathematical formulation, the generality of his concepts and procedures (he has, for example, no need to distinguish a special ground line and horizon line), and the completely new concise synthesis which he did achieve of procedures not all of which were original with him.

The first of the three specific items which will be discussed here is his construction for the perspective of a triangle ABC (see Figure 20 of Taylor's Plate 7 as reproduced in our Figure 2). Consider the plane of the drawing to represent the picture plane with two other planes rotated into coincidence with it. Below ED , the intersection of the original plane and the picture, is the original plane itself, containing ABC (to be thought of as "behind" the picture from the viewpoint of the observer) rotated about ED into the picture. The plane through the eye parallel to the ground plane has been rotated upward about FH . Above FH then is O , the eye point. Extend AB to meet ED in D , its "intersection". Draw a line through O parallel to AB to meet FH in F , the vanishing point of AB . FD is then the indefinite perspective of AB , i.e., the perspective of AB produced. Join O to A and B . The intersections of these lines with FD determine perspective points a and b . A similar determination of c (c could also be located as the intersection of EH and IG) would

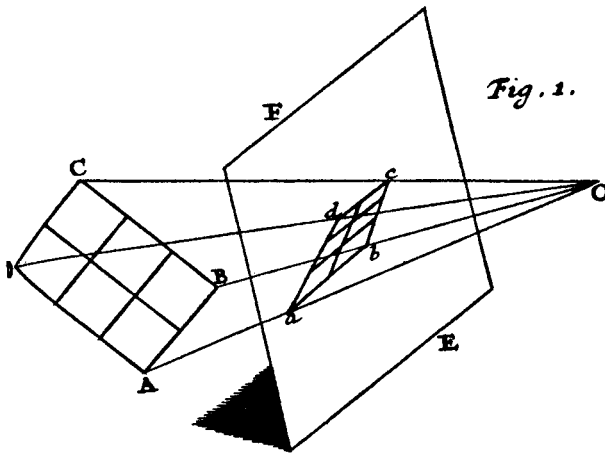


Fig. 1.

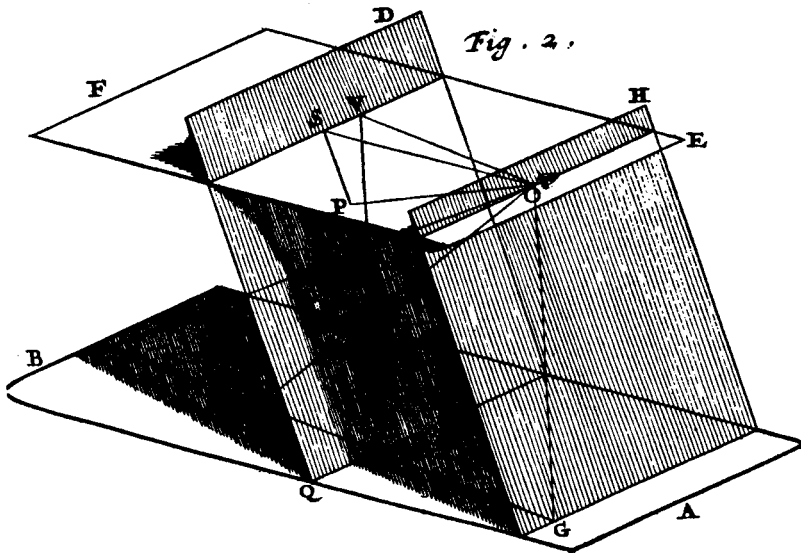


Fig. 2.

Figure 1.

give a diagram in which the corresponding sides of the two triangles meet on ED and the lines joining corresponding vertices concur in O . Although the Desargues triangle theorem is neither mentioned nor stated, note how completely it is implicit in this construction and the accompanying diagram [8]. Both the problem and the diagram were modified in the second edition and the relationship, though still implicit, became less obvious.

Also in the 1715 edition but omitted in the second edition is the problem of finding the perspective of the shadow of a triangle on a plane. Not only does this associate with the three dimensional case of the Desargues theorem, but of particular interest is Taylor's second solution of the problem which is, as he terms it, by putting the rules of perspective in per-

spective. In this same vein he elsewhere gives constructions for such things as the vanishing point of lines perpendicular to a given plane for the specific purpose of making it possible to draw directly in perspective without first having an orthogonal projection. In this Taylor anticipated Lambert who took this as one of the major objectives of his *Freye Perspective* (1759). Taylor's work with such problems led him to make repeated use in the 1715 edition of the idea of associating infinitely distant intersections with parallel lines.

A second construction which is both unique to the 1715 edition and which has for its purpose the construction of drawings directly in perspective is Taylor's solution of the problem of completing the construction of the perspective of a circle, given the

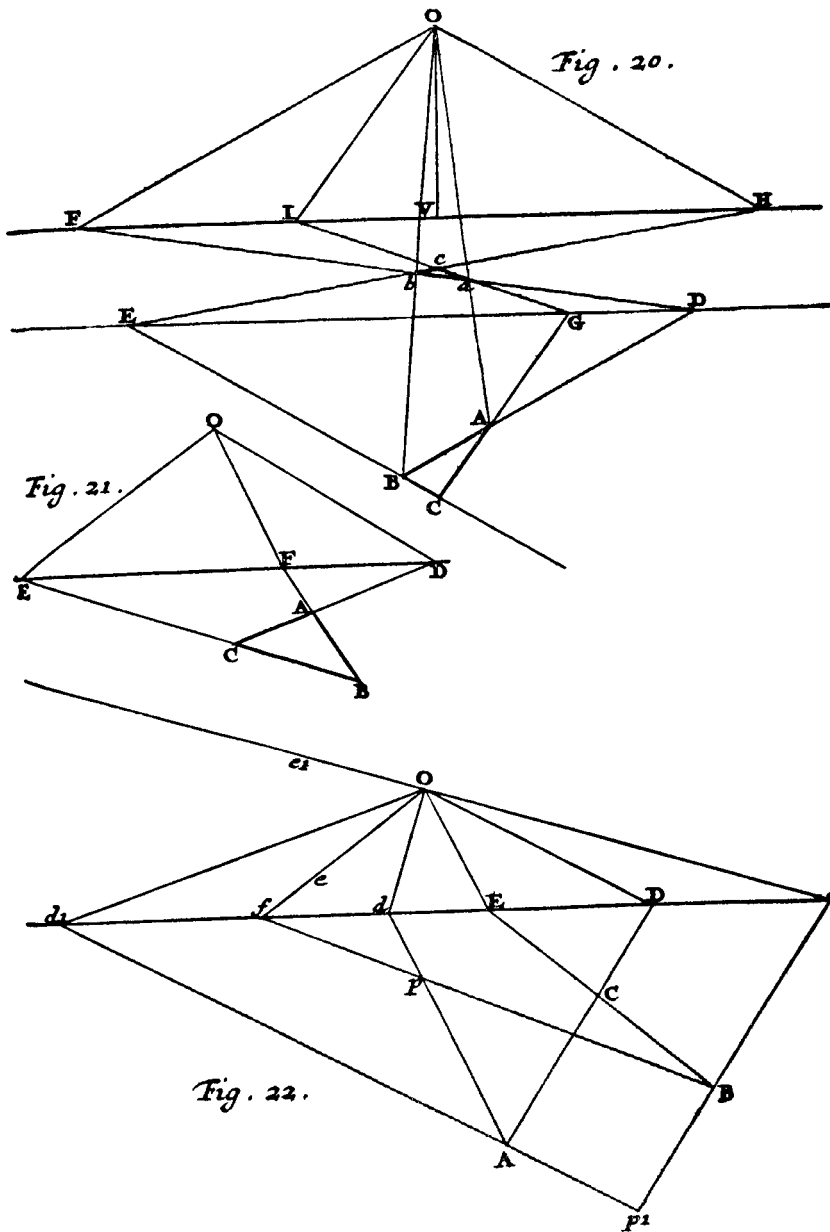


Figure 2.

perspective of its center and of one of its points. The diagram for this is to be found in "Fig. 21" of Taylor's Plate 7 which is our Figure 2. C is the perspective of the center of a circle, A the perspective of a point on the circle, ED the vanishing line of its plane, and O the eye rotated into the picture. CA then represents a radius. Draw any line through C to meet the vanishing line in E and extend CA to intersect it in D . Bisect angle EOD to locate point F on ED . The join of F and A meets EC in B ,

another point of the perspective circle. Taylor's reasoning was based on the fact that since the angles at A and B are perspectives of equal angles then CA and CB are perspectives of the sides of an isosceles triangle and hence are the perspectives of equal lines. CB must then represent a radius, and B is the perspective of a point on the circle. This is another example of Taylor's thinking and drawing directly in perspective. It is also interesting to note that if the construction were extended to determine the second

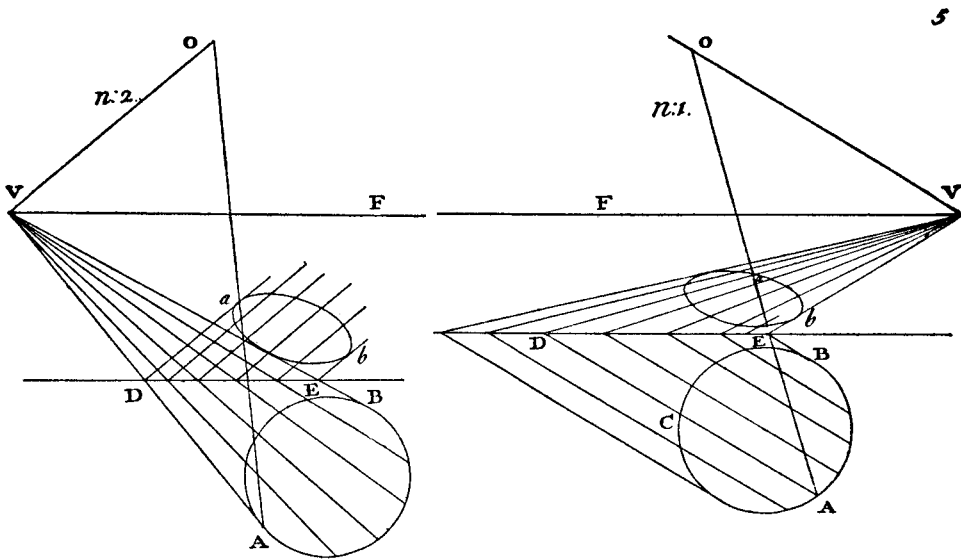


Figure 3.

point on each radius by bisecting the supplement of EOD we would have an harmonic set of points on ED , and further that C and ED are pole and polar with respect to the conic which represents the circle.

The first explicit use which the author has found of the terms pole and polar in perspective is in the work of Cousinery in 1828. However, John Hamilton, one of Taylor's followers, had also read LaHire's *Sectiones Conicae* (1685) in which much use is made of harmonic sets. Book III of Hamilton's *Stereography or a Compleat Body of Perspective* (1738) makes extensive use of harmonic sets and some use of theorems on poles and polars although without using the latter terms.

Our Figure 3 shows, for contrast with the above, the two constructions for the perspective of a circle which were given as Figure 13 in the 1719 edition of Taylor's book. They are more conventional, use the orthogonal projection of the original circle, and are described in much more detail in the text.

Taylor gives no proof or explanation of the third unique construction which is here presented from the 1715 edition. The construction we refer to is "n:2." in Figure 32 of Taylor's Plate 12 which is shown here as Figure 4. Both "n:1." and "n:2." are constructions for a line through a given point and the inaccessible intersection of two other given lines. Today, "n:2." would be regarded as an application of harmonic sets related to complete quadrilaterals. Knowing that he did use both the idea that lines meeting on a vanishing line are parallel and its con-

verse, we can guess that Taylor might have proven it quickly and easily by thinking in a "perspective" geometry where $ABCD$ and $abcd$ would be parallelograms rather than in a "Euclidean" geometry. In any case, Taylor was the first writer on perspective to treat this problem.

The only original work on perspective printed in England prior to Taylor's was Humphrey Ditton's *A Treatise of Perspective* of 1712, which deserves more note than it has had in the past but which is not comparable to Taylor's *Linear Perspective* in generality or originality. Following Taylor in England only Hamilton showed much originality while on the continent Lambert's work was outstanding in this century. Another feature which, though first met in Guido Ubaldo Monte's *Perspectivae Librix Sex* (Pisa, 1600), was developed by Taylor and then carried much farther by Lambert was the solution of the inverse problem of perspective. This problem, namely, given data about a perspective drawing to draw inferences about the original, is basic in the modern science of photogrammetry.

This discussion represents only a portion of a complete study which the author has made of the development of the mathematical theory of perspective. It shows that Brook Taylor contributed a mathematically clear, concise, and logical, but abstract, formulation of extraordinary generality, including some treatment of the inverse problem. The editions, translations, and sequels to his work noted here extended his influence beyond both his homeland and his

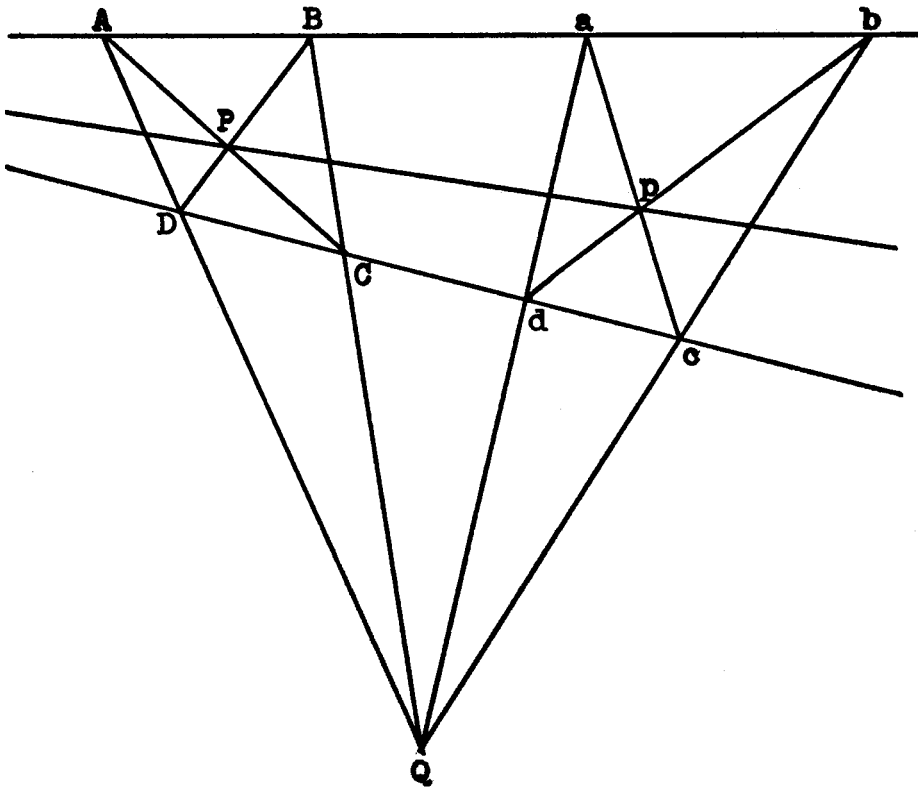


Figure 4.

chronological period.

In conclusion, we note for those who might wonder at the interest of Taylor in this subject that not only is this interest consistent with the mathematical and cultural interests of the time (Desargues, Stevin, Ozanam, were earlier mathematicians who wrote on this topic), but also that Taylor grew up in a home where music and art were popular diversions. According to his grandson, Taylor himself in his painting “favored landscapes and water colors. They have a force of color, a freedom of touch, a varied disposition of planes of distance, and a learned use of aerial as well as linear perspective which all professional men who have seen these paintings have admired” [9].

Notes

1. Max Steck, *Johann Heinrich Lambert Schriften zur Perspektive*. (Berlin: 1943), p. 48 lists Jacquier’s French translation of Taylor’s work, as among the books in Lambert’s library and adds parenthetically that it was “von Lambert im II Teil des Hauptwerkes benutzt.”
2. Since the writer found no other at all complete enumer-

ation of these works, it seems appropriate to preserve this data in detail for future reference. The books referred to are:

Editions:

- Brook Taylor, *Linear Perspective*, London: 1715.
 Brook Taylor, *New Principles of Linear Perspective*, London: 1719.
 Brook Taylor, *New Principles of Linear Perspective*, third edition corrected by J. Colson. London: 1749.
 Brook Taylor, *Method of Perspective*, 1766. The *Dictionary of National Biography* (London: 1899 LIX, p. 359) lists this under Isaac Ware who, it says, prepared the edition.
 Brook Taylor, *New Principles of Linear Perspective: The fourth edition, revised*, London: 1811.

Translations:

- Francois Jacquier, *Elementi di prospettiva secondo li principii di Brook Taylor, con varie aggiunte*, Roma: 1753.
 Antoine Rivoire, *Nouveaux principes de la perspective lineaire, traduction de deux ouvrages, l’un Anglois, due Docteur Brook Taylor, l’autre Latin, de M. Patrice Murdoch*, Amsterdam: 1759.

Jacopo Stellini, *Opere varie*, Padova: 1781. Contains in volume II Taylor's "Nuovi principij della prospettiva lineare" according to Pietro Riccardi in his *Biblioteca Matematica Italiana*.

Disciples:

John Hamilton, *Stereography or a compleat body of perspective*, London: 1738, 1740, 1748.

John Joshua Kirby, *Dr. Brook Taylor's method of perspective made easy both in theory and practice*, Ipswich: 1754, 1755; London: 1765, 1768.

John Joshua Kirby, *The perspective of architecture — deduced from the principles of Dr. Brook Taylor*, London: 1761.

John Joshua Kirby, *Dr. Brook Taylor's method of perspective, compared with examples lately published on this subject, as Sirigatti's by Isaac Ware*, London: 1767.

Daniel Fournier, *A treatise of the theory and practice of perspective. Wherein the principles — laid down by Dr. Brook Taylor are explained by moveable schemes*, London: 1761, 1762, 1763, 1764.

Joseph Highmore, *Practice of perspective on the principles of Dr. Brook Taylor*, London: 1763.

Thomas Malton, *A compleat treatise on perspective in theory and practice on the true principles of Dr. Brook Taylor*, London: 1775, 1776, 1779.

Thomas Malton, *An appendix or second part to the compleat treatise on perspective containing a brief history of perspective*, London: 1783.

James Malton, *The young painter's maulstick; being a practical treatise on perspective; — with the theoretic principles of B. Taylor*, London: 1800.

Edward Edwards, *A practical treatise of perspective on the principles of Dr. Brook Taylor*, London: 1803.

Joseph Jopling, *Taylor's principles of linear perspective, new edition with additions by Joseph Jopling*, London: 1835.

George Blacker, *John Heywood's second grade perspective adapted from Dr. Brook Taylor*, Manchester: 1885–88.

3. *Contemplatio Philosophica: A Posthumous Work of the Late Brook Taylor, L.L.D., F.R.S. some time Secretary of the Royal Society. To which is prefixed a life of the author by his grandson, Sir William Young, Bart. F.R.S., A.S.S.* (London: Printed by W. Bulmer and Co., 1793), p. 29. The title page of this book bears the printed note *Not Published*. The book also includes some letters to and from Taylor to which we will refer later.
4. "Epistola Pro Eminente Mathematico Dn. Johanne Bernoullio, contra quendam ex Anglia antagonistam scripta" *Acta Eruditorum*, (July, 1716), pp. 296–315. The article preceding this one in *Acta* was "Methodus Incrementorum Directa & Inversa; Autore Broock (sic!) Taylor, L.L.D. & Regiae Societatis Secretario," a summary of the book with comments, references to Leibniz and his procedures and to Collins' *Commercium Epistolicum*. This "review" was probably written by Leibniz himself according to Heinrich Aucter, *Brook Taylor der Mathematiker und Philosoph*, (Wurzburg: Konrad Tritsch, 1937), p. 79.

The Taylor-Bernoulli dispute as it appeared in *Acta* and the *Philosophical Transactions* is somewhat expanded in details and clarified by the letters printed by Young in the *Contemplatio Philosophica* and in Aucter, *op. cit.*

Taylor wrote on February 5, 1719 to Count Raymond de Montmort in reply to a letter from Bernoulli which Montmort had forwarded, "For if the book be so very obscure, as he says it is, that the best artists, those already acquainted with the subjects, cannot well understand it." This may be the source for Young's quotation. Taylor, however, seems to have been referring to his *Methodus* rather than his work on perspective.
5. Institut de France, Académie des Sciences, *Procès-Verbaux des Seances de l'Académie*, Tome II, An VIII–XI (1800–1804), 1912, pp. 360 ff.
6. Julian L. Coolidge, *A History of Geometrical Methods*, (Oxford, 1940), p. 108.
7. Gino Loria, *Storia della Geometria Descrittiva*, (Milano, 1921), pp. 43–51.
8. The copy of Taylor's book used originally in this study is in the Rare Book Room of the University of Michigan. The author, happening recently to have purchased a copy for himself, was startled to find lines *OaA* and *ObB* in his copy to have been drawn in with ordinary pen and ink after printing. Further comparison of the two copies showed that a number of corrections to the plates were made during the printing process, appearing inked in the author's copy and printed in the library's copy. It should be remarked that practically all of the original works cited are to be found in the University of Michigan's collection built up by Professor L. C. Karpinski whose suggestions and advice aided significantly in the study of which this paper is a partial report.
9. Sir W. Young, *op. cit.*, pp. 28–29.

Was Newton's Calculus a Dead End? The Continental Influence of Maclaurin's Treatise of Fluxions

JUDITH GRABINER

American Mathematical Monthly 104 (1997), 393–410

1 Introduction

Eighteenth-century Scotland was an internationally-recognized center of knowledge, “a modern Athens in the eyes of an enlightened world.” [74, p. 40] [81] The importance of science, of the city of Edinburgh, and of the universities in the Scottish Enlightenment has often been recounted. Yet a key figure, Colin Maclaurin (1698–1746), has not been highly rated. It has become a commonplace not only that Maclaurin did little to advance the calculus, but that he did much to retard mathematics in Britain—although he had (fortunately) no influence on the Continent. Standard histories have viewed Maclaurin's major mathematical work, the two-volume *Treatise of Fluxions* of 1742, as an unread monument to ancient geometry and as a roadblock to progress in analysis. Nowadays, few people read the *Treatise of Fluxions*. Much of the literature on the history of the calculus in the eighteenth and nineteenth centuries implies that few people read it in 1742 either, and that it marked the end—the dead end—of the Newtonian tradition in calculus. [9, p. 235], [49, p. 429], [10, p. 187], [11, pp. 228–9], [43, pp. 246–7], [42, p. 78], [64, p. 144]

But can this all be true? Could nobody on the Continent have cared to read the major work of the leading mathematician in eighteenth-century Scotland? Or, if the work was read, could it truly have been “of little use for the researcher” [42, p. 78] and have had “no influence on the development of mathematics”? [64, p. 144]

We will show that Maclaurin's *Treatise of Flux-*

ions did develop important ideas and techniques and that it did influence the mainstream of mathematics. The Newtonian tradition in calculus did not come to an end in Maclaurin's Britain. Instead, Maclaurin's *Treatise* served to transmit Newtonian ideas in calculus, improved and expanded, to the Continent. We will look at what these ideas were, what Maclaurin did with them, and what happened to this work afterwards. Then, we will ask what by then should be an interesting question: why has Maclaurin's role been so consistently underrated? These questions will involve general matters of history and historical writing as well as the development of mathematics, and will illustrate the inseparability of the external and internal approaches in understanding the history of science.

2 The standard picture

Let us begin by reviewing the standard story about Maclaurin and his *Treatise of Fluxions*. The calculus was invented independently by Newton and Leibniz in the late seventeenth century. Newton and Leibniz developed general concepts—differential and integral for Leibniz, fluxion and fluent for Newton—and devised notation that made it easy to use these concepts. Also, they found and proved what we now call the Fundamental Theorem of Calculus, which related the two main concepts. Last but not least, they successfully applied their ideas and techniques to a wide range of important problems. [9, p. 299] It was not until the nineteenth century, however, that the basic concepts were given a rigorous foundation.

In 1734 George Berkeley, later Bishop of Cloyne, attacked the logical validity of the calculus as part of his general assault on Newtonianism. [12, p. 213] Berkeley's criticisms of the rigor of the calculus were witty, unkind, and—with respect to the mathematical practices he was criticizing—essentially correct. [6, v. 4, pp. 65–102] [38, pp. 33–34] [82, pp. 332–338] Maclaurin's *Treatise* was supposedly intended to refute Berkeley by showing that Newton's calculus was rigorous because it could be reduced to the methods of Greek geometry. [10, pp. 181–2, 187] [9, pp. 233, 235] Maclaurin himself said in his preface that he began the book to answer Berkeley's attack, [63, p. i] and also to rebut Berkeley's accusation that mathematicians were hostile to religion. [78, p. 50]

The majority of Maclaurin's *Treatise* is contained in its first Book, which is called "The Elements of the Method of Fluxions, Demonstrated after the Manner of the Ancient Geometricians." That title certainly sounds as though it looks backward to the Greeks, not forward to modern analysis. And the text is full of words—lots of words. So much time is spent on preliminaries that it is not until page 162 that he can show that the fluxion of ay is a times the fluxion of y . Florian Cajori, whose writings have helped spread the standard story, compared Maclaurin to the German poet Klopstock who, Cajori said, was praised by all, read by none. [10, p. 188] While British mathematicians, bogged down with geometric baggage, studied and revered the work and notation of Newton and argued with Berkeley over foundations, Continental mathematicians went onward and upward analytically with the calculus of Leibniz. The powerful analytic results and techniques in eighteenth-century Continental mathematics were all that mathematicians like Cauchy, Riemann, and Weierstrass needed for their nineteenth-century analysis with its even greater power, together with its improved rigor and generality. [9, ch. 7] [49, p. 948] This story became so well known that it was cited by the literary critic Matthew Arnold, who wrote, "The man of genius [Newton] was continued by . . . completely powerless and obscure followers . . . The man of intelligence [Leibniz] was continued by successors like Bernoulli, Euler, Lagrange, and Laplace—the greatest names in modern mathematics." [1, p. 54; cited by [61, p. 15]]

Now since I myself have contributed to the standard story, especially in delineating the links among Euler, Lagrange, and Cauchy, [38, chs. 3–6] I have a good deal of sympathy for it, but I now think that it

must be modified. Maclaurin's *Treatise of Fluxions* is an important link between the calculus of Newton and Continental analysis, and Maclaurin contributed to key developments in the mathematics of his contemporaries. Let us examine the evidence for this statement.

3 The nature of Maclaurin's *Treatise of Fluxions*

Why—the standard story notwithstanding—might Maclaurin's *Treatise of Fluxions* have been able to transmit Newtonian calculus, improved and expanded, to the Continent? First, because the *Treatise of Fluxions* is not just one "Book," but two. While Book I is largely, though not entirely, geometric, Book II has a different agenda. Its title is "On the *Computations* in the Method of Fluxions." [my italics] Maclaurin began Book II by championing the power of symbolic notation in mathematics. [63, pp. 575–576] He explained, as Leibniz before him and Lagrange after him would agree, that the usefulness of symbolic notation arises from its generality. So, Maclaurin continued, it is important to demonstrate the rules of fluxions once again, this time from a more algebraic point of view. Maclaurin's appreciation of the algorithmic power of algebraic and calculus notation expresses a common eighteenth-century theme, one developed further by Euler and Lagrange in their pursuit of pure analysis detached from any kind of geometric intuition. To be sure, Maclaurin, unlike Euler and Lagrange, did not wish to detach the calculus from geometry. Nonetheless, Maclaurin's second Book in fact, as well as in rhetoric, has an algorithmic character, and most of its results may be read independently of their geometric underpinnings, even if Maclaurin did not so intend. (In his Preface to Book I, he even urged readers to look at Book II before the harder parts of Book I.) [63, p. iii] The *Treatise of Fluxions*, then, was not foreign to the Continental point of view, and may have been written in part with a Continental audience in mind.

Nor was this algebraic character a secret open only to the reader of English. There was a French translation in 1749 by the Jesuit R. P. Pézénas, including an extensive table of contents. [62] Lagrange, among others, seems to have used this French edition (since he cited it by the French title [58, p. 17] though he cited other English works in English [58, p. 18]). Pézénas' translation, moreover, was neither isolated nor idiosyncratic, but part of the activity

of a network of Jesuits interested in mathematics and mathematical physics, especially work in English, with Maclaurin one of the authors of interest to them. [84, pp. 33, 221, 278, 517, 655] For instance, Pézénas himself translated other English works, including those by Desaguliers, Gardiner's logarithmic tables, and Seth Ward's *Young Mathematician's Guide*. [83, pp. 571–2] Thus there was a well-worn path connecting English-language work with interested Continental readers. Furthermore, the two-fold character of the *Treatise of Fluxions* was noted, with special praise for Book II's treatment of series, by Silvestre-François Lacroix in the historical introduction to the second edition of his highly influential three-volume calculus textbook. [52, p. xxvii] Unfortunately, though, recognition of the two-fold character has been absent from the literature almost completely from Lacroix's time until the recent work by Sageng and Guicciardini. [42] [78] We shall address the reasons for this neglect in due course.

4 The social context: The Scottish Enlightenment

Another reason for doubting the standard picture comes from the social context of Maclaurin's career. Eighteenth-century Scotland, Maclaurin's home, was anything but an intellectual backwater. It was full of first-rate thinkers who energetically pursued science and philosophy and whose work was known and respected throughout Europe. One would expect Scotland's leading mathematician to share these connections and this international renown, and he did.

Although Scotland had been deprived of its independent national government by the Act of Union of 1707, it still retained, besides its independent legal system and its prevailing religion, its own educational system. The strength and energy of Scottish higher education in Maclaurin's time is owed in large part to the Scottish ruling classes, landowners and merchants alike, who saw science, mathematics, and philosophy as keys to what they called the "improvement" of their yet underdeveloped nation. [65, p. 254] [80, pp. 7–8, 10–11] [17, pp. 127, 132–3] Eighteenth-century Scotland, with one-tenth the population of England, had four major universities to England's two. [80, p. 116] Maclaurin, when he wrote the *Treatise of Fluxions*, was Professor of Mathematics at the University of Edinburgh. Edinburgh was about to become the heart of the Scottish Enlightenment, and Maclaurin until his death in

1746 was a leading figure in that city's cultural life.

Mathematics played a major role in the Scottish university curriculum. This was in part for engineers; Scottish military engineers were highly in demand even on the Continent. [17, p. 125] Maclaurin himself was actively interested in the applications of mathematics, and just before his untimely death had planned to write a book on the subject. [36] [68, p. xix] In addition, mathematics and Newtonian physics were part of the course of study for prospective clergymen. [80, p. 20] The influential "Moderate" party in the Church of Scotland appreciated the Newtonian reconciliation of science and religion. [16, pp. 53, 57]

Maclaurin's position in Edinburgh's cultural life was not just that of a technically competent mathematician. For instance, he was part of the Rankenian society, which met at Ranken's Tavern in Edinburgh to discuss such things as the philosophy of Bishop Berkeley; the society introduced Berkeley's philosophy to the Scottish university curriculum. [24, p. 222] [17, p. 133] [65, p. 197] Maclaurin and his physician friend Alexander Monro were the founders and moving spirits of the Edinburgh Philosophical Society. [65, p. 198] With Newton's encouragement, Maclaurin had become the chief spokesman in Scotland for the new Newtonian physics. His posthumously published book, *An Account of Sir Isaac Newton's Philosophical Discoveries*, was based on material Maclaurin used in his classes at Edinburgh, and the book was of great interest to philosophers. [24, p. 137] That book became well known on the Continent. It was translated into French almost as soon as it appeared, by Louis-Anne Lavoisier in 1749, and the first part appeared in Italian in Venice in 1762.

Another branch of Scottish science, namely medicine, also had many links with the Continent and was highly regarded there. Medical students went back and forth between Scotland, Holland, and France. [17, p. 135] [80, p. 7]

The best-known figures of eighteenth-century Scotland had major interactions with, and influence upon, Continental science and philosophy. [39] [81] Let it suffice to mention the names of four: the philosopher David Hume, who was a student at Edinburgh in Maclaurin's time; the geologist James Hutton, who attended and admired Maclaurin's lectures; [34, pp. 577–8] and, a bit after Maclaurin's time but still subject to his influence on Scottish higher education, the chemist Joseph Black and the economic and political philosopher Adam Smith.

Maclaurin himself had twice won prizes from the Académie des Sciences in Paris, once in 1724 for a memoir on percussion, and then in 1740 (dividing the prize with Daniel Bernoulli, P. Antoine Cavalieri, and Leonhard Euler) for a memoir on the tides. [79, p. 611] [39, pp. 400–401]

Scotland in the eighteenth century nurtured first-rate intellectual work on mathematics, philosophy, science, medicine, and engineering, and did it all as part of a general European culture. [39, p. 412] [81, *passim*] The *Treatise of Fluxions* was the major mathematical work of a Scottish mathematician of considerable reputation on the Continent, a major work philosophically attuned to the enormously influential Newtonian physics and the Continentally popular algebraic symbolism. Such a work would certainly be of interest to Continental thinkers. Social considerations may not suffice to determine mathematical ideas, but they certainly affect the mathematician's ability to make a living, to get research support, and to promote contact and communication with other mathematicians and scientists at home and abroad. And so it was with Maclaurin.

5 Maclaurin's Continental reputation

An even better reason for not accepting the traditional view of Maclaurin is that his work demonstrably was read in the eighteenth century, and was read by the big names of Continental mathematics. He had a Continental acquaintance through travel and correspondence. Even before the *Treatise of Fluxions*, his reputation had been enhanced by his Académie prizes and by his books on geometry. He was thus a respected member of an international network of mathematicians with interests in a wide range of subjects, and the publication of the *Treatise of Fluxions* was eagerly anticipated on the Continent.

The *Treatise of Fluxions* of 1742 was Maclaurin's major work on analysis, incorporating and somewhat dwarfing what he had done earlier. It contains an exposition of the calculus, with old results explained and many new results introduced and proved. Maclaurin seems to have included almost everything he had done in analysis and its applications to Newtonian physics. In particular, the findings of his Paris prize paper on the tides were included and expanded. His other papers, the posthumous and relatively elementary Algebra, and his works on geometry as such—though highly regarded—do not

concern us here, but his Continental reputation was enhanced by these as well.

Let us turn now to some specific evidence for the Continental reputation of Maclaurin's major work. In 1741, Euler wrote to Clairaut that, though he had not yet seen the Paris prize papers on the tides, "from Mr. Maclaurin I expect only excellent ideas." [47, p. 87] Euler added that he had heard from England (presumably from his correspondent James Stirling) that Maclaurin was bringing out a book on "differential calculus," and asked Clairaut to keep him posted about this. In turn, Clairaut asked Maclaurin later in 1741 about his plans for the book, [66, p. 348] which Clairaut wanted to see before publishing his own work on the shape of the earth. [47, p. 110] Euler did get the *Treatise of Fluxions*, and read enough of it quickly to praise it in a letter to Goldbach in 1743. [48, p. 179] Jean d'Alembert, in his *Traité de dynamique* of 1743, [22, sec. 37, n.] praised the rigor brought to calculus by the *Treatise of Fluxions*. D'Alembert's most recent biographer, Thomas Hankins, argues that Maclaurin's *Treatise*, appearing at this time, helped persuade d'Alembert that gravity could best be described as a continuous acceleration rather than a series of infinitesimal leaps. [44, p. 167] D'Alembert's general approach to the foundations of the calculus in terms of limits clearly was influenced by Newton's and Maclaurin's championing of limits over infinitesimals, in particular by Maclaurin's clear description of limits in one of the parts of his *Treatise of Fluxions* that explicitly responds to Berkeley's objections (and which incidentally may be the first explicit description of the tangent as the limit of secant lines; see Section 7 below). [44, p. 23] [63, pp. 422–3] Lagrange in his *Analytical Mechanics* [55, p. 243] said that Maclaurin, in the *Treatise of Fluxions*, was the first to treat Newton's laws of motion in the language of the calculus in a coordinate system fixed in space. Though C. Truesdell [80, pp. 250–3] has shown that Lagrange was wrong because Johann Bernoulli and Euler were ahead of Maclaurin on this, the fact that Lagrange believed this is one more piece of evidence of the Continental reputation of Maclaurin as mathematician and physicist.

6 Maclaurin's mathematics and its importance

The previous points show that Maclaurin could have been influential, but not that he was. Five examples

will reveal both the nature of Maclaurin's techniques and the scope of his influence: a special case of the Fundamental Theorem of Calculus; Maclaurin's treatment of maxima and minima for functions of one variable; the attraction of spheroids; what is now called the Euler-Maclaurin summation formula; and elliptic integrals.

a. Key methods in the calculus Two methods were central to the study of real-variable calculus in the eighteenth and nineteenth centuries. One of these is studying real-valued functions by means of power-series representations. This tradition is normally thought first to flower with Euler; it is then most closely associated with Lagrange, and, later for complex variables, with Weierstrass. The second such method is that of basing the foundations of the calculus on the algebra of inequalities—what we now call delta-epsilon proof techniques—and using algebraic inequalities to prove the major results of the calculus; this tradition is most closely associated with the work of Cauchy in the 1820's. I have traced these traditions back to Lagrange and Euler in my work on the origins of Cauchy's calculus. [38, chs. 3–6] It is surprising, at least if one accepts the standard picture of the history of the calculus, that both of these methods—studying functions by power series, basing foundations on inequalities—were materially advanced by Maclaurin in the *Treatise of Fluxions*. It is especially striking that the importance of Maclaurin's work on series—work based, it is well to remember, on Newton's use of infinite series—was recognized and praised in 1810 by Lacroix, who also linked it with the series-based calculus of Lagrange. [52, p. xxxiii]

Maclaurin skillfully used algebraic inequalities in his proof of a special case of the Fundamental Theorem of Calculus. He showed, for a particular function, that if one takes the fluxion of the area under the curve whose equation is $y = f(x)$, one gets the function $f(x)$. In his proof, Maclaurin adapted the intuition underlying Newton's argument for this fact in *De Analysis* [69]—that the rate of change of the area under a curve is measured by the height of the curve—but Maclaurin's proof is more rigorous. Although Maclaurin's argument proceeds algebraically, the concepts involved resemble those of the Greek “method of exhaustion” (more precisely termed by Dijksterhuis “indirect passage to the limit”). [26, p. 130] A key step in this Greek work is first to assume that two equal areas or expressions for areas are *unequal*, and then to argue to a contradiction by

using inequalities that hold among various rectilinear areas. Newton in the *Principia* had based proofs of new results about areas and curves on methods akin to those of the Greeks. Maclaurin carried this much further. It was Maclaurin's “conservative” allegiance to Archimedean geometric methods that led him to buttress the *kinematic* intuition of Newton's calculus with *algebraic* inequality proofs.

What Maclaurin proved in the example under discussion is that, if the area under a curve up to x is given by x^n , the ordinate of the curve must be $y = nx^{n-1}$, which is known to be the fluxion of x^n . [63, pp. 752–754] Maclaurin's diagram for this is much like the one Newton gave in the *De Analysis*. [69, pp. 3–4] Maclaurin began by saying that, since x and y increase together, the following inequality holds between the areas shown:

$$x^n - (x - h)^n < yh < (x + h)^n - x^n. \quad (1)$$

(Maclaurin gave this inequality verbally; I have supplied the “<” signs; also, I use “ h ” for the increment where Maclaurin used “ o ”). Now Maclaurin recalled an algebraic identity he had proved earlier: [63, p. 583; inequality notation added]

If $E < F$, then

$$nF^{n-1}(E - F) < E^n - F^n < nE^{n-1}(E - F). \quad (2)$$

(It may strike the modern reader that, since nx^{n-1} is the derivative of x^n , this second inequality is a special case of the mean-value theorem for derivatives. I shall return to this point later.)

Now, letting $x - h$ play the role of F and x play the role of E , $E - F$ is h and the first inequality in (2) yields

$$n(x - h)^{n-1}h < x^n - (x - h)^n.$$

Similarly, if $F = x$ and $E = x + h$, then $E - F = h$ and the second inequality in (2) becomes

$$(x + h)^n - x^n < n(x + h)^{n-1}h.$$

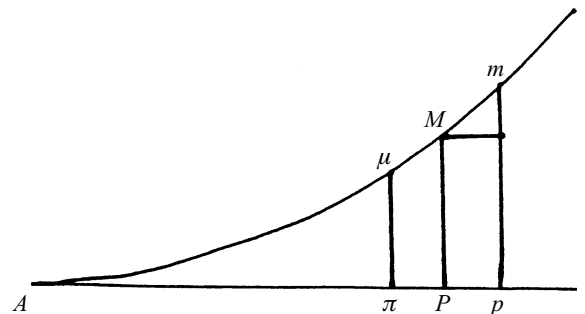


Figure 1.

Combining these with inequality (1) about the areas, Maclaurin obtained

$$n(x-h)^{n-1}h < yh < n(x+h)^{n-1}h.$$

Dividing by h produces

$$n(x-h)^{n-1} < y < n(x+h)^{n-1}. \quad (3)$$

Recall that, given that the area was x^n , Maclaurin was seeking an expression for y , the fluxion of that area. A modern reader, having reached the inequality (3), might stop, perhaps saying “let h go to zero, so that y becomes nx^{n-1} ,” or perhaps justifying the conclusion by appealing to the delta-epsilon characterization of limit. What Maclaurin did instead was what Archimedes might have done, a double *reductio ad absurdum*. But what Archimedes might have done geometrically and verbally, Maclaurin did algebraically. He assumed first that y is *not* equal to nx^{n-1} . Then, he said, it must be equal to $nx^{n-1} + r$ for some r . First, he considered the case when this r was positive. This will lead to a contradiction if h is chosen so that $y = n(x+h)^{n-1}$, since, he observed, inequality (3) will be violated when $h = (x^{n-1} + r/n)^{1/(n-1)}$. Similarly, he calculated the h that produces a contradiction when r is assumed to be negative. Thus there can be no such r , and $y = nx^{n-1}$. [63, p. 753]

Maclaurin introduced this proof by saying something surprising for a *Treatise of Fluxions*: that the use of the inequalities makes the demonstration of the value of y “independent of the notion of a fluxion.” [63, p. 752] (Of course one would need the notion of fluxion to interpret y as the fluxion of the area function x^n , but the proof itself is algebraic.) This proof was presumably part of his agenda in writing the more algebraic Book II of the *Treatise* for an audience on the Continent, where fluxions were suspect as involving the idea of motion. Later Lagrange, in seeking his purely algebraic foundation for the calculus, explicitly said he wanted to free the calculus from fluxions and what he called the “foreign idea” of motion. It is thus striking that Lagrange’s *Théorie des fonctions analytiques* (1797) gives a more general version of the kind of argument Maclaurin had given, applying to any increasing function that satisfies the geometric inequality expressed in (1). In place of the algebraic inequality (2), Lagrange used the mean-value theorem. [58, pp. 238–9] [38, pp. 156–158] The similarity of the two arguments does not prove influence, of course, but it certainly demonstrates that Maclaurin’s work, which

we know Lagrange read (e.g., [58, p. 17]), uses the algebra of inequalities in a way consistent with that used by Lagrange and his successors.

Maclaurin’s argument exemplifies the way his *Treatise* reconciles the old and the new. The double *reductio ad absurdum* reflects his Archimedean agenda. Treating the area as generated by a moving vertical line, and then searching for the relationship between the area and its fluxion, are Newtonian. Maclaurin did not have a general proof of the Fundamental Theorem in this argument, but relied on an inequality based on the specific properties of a specific function. Nonetheless, he had the precise bounding inequalities for the area function used later by Lagrange, and he used an algebraic inequality proof in a manner that would not disgrace a nineteenth-century analyst.

Inequality-based arguments in the calculus as used by Lagrange and Cauchy owe a lot to the eighteenth-century study of algebraic approximations, and it once seemed to me that this was their origin. But the algebra of inequalities as used in Continental analysis, especially in d’Alembert’s pioneering treatment of the tangent as the limit of secants in the article “Différentiel” in the *Encyclopédie*, [19] must owe something also to Maclaurin’s translation of Archimedean geometry into algebraic dress to justify results in calculus. Throughout the eighteenth century, practitioners of the limit tradition on the Continent use inequalities; a clear line of influence connects Maclaurin’s admirer d’Alembert, Simon L’Huilier (who was a foreign member of the Royal Society), the textbook treatment of limits by Lacroix, and, finally, Cauchy. [38, pp. 80–87]

Now let us turn to some of Maclaurin’s work on series. There is, of course, the Maclaurin series, that is, the Taylor series expanded around zero. This result Maclaurin himself credited to Taylor, and it was known earlier to Newton and Gregory. It was called the Maclaurin series by John. F. W. Herschel, Charles Babbage, and George Peacock in 1816 [51, pp. 620–21] and by Cauchy in 1823. [14, p. 257] Since it was obvious that Maclaurin had not invented it, the attribution shows appreciation by these later mathematicians for the way Maclaurin used the series to study functions. A key application is Maclaurin’s characterization of maxima, minima, and points of inflection of an infinitely differentiable function by means of its successive derivatives. When the first derivative at a point is zero, there is a maximum if the second derivative is negative there, a minimum if it is positive. If the second derivative is zero, one

looks at higher derivatives to tell whether the point is a maximum, minimum, or point of inflection. These results can be proved by looking at the Taylor series of the function near the point in question, and arguing on the basis of the inequalities expressed in the definition of maximum and minimum. For instance (in modern [Lagrangian] notation), if $f(x)$ is a maximum, then

$$\begin{aligned} f(x) &> f(x+h) \\ &= f(x) + hf'(x) + h^2/2!f''(x) + \dots \end{aligned} \quad (4)$$

and

$$\begin{aligned} f(x) &> f(x-h) \\ &= f(x) - hf'(x) + h^2/2!f''(x) - \dots \end{aligned}$$

if h is small. If the derivatives are bounded, and if h is taken sufficiently small so that the term in h dominates the rest, the inequalities (4) can both hold only if $f'(x) = 0$. If $f'(x) = 0$, then the h^2 term dominates, and the inequalities (4) hold only if $f''(x)$ is negative. And so on.

I have traced Cauchy's use of this technique back to Lagrange, and from Lagrange back to Euler. [38, pp. 117–118] [37, pp. 157–159] [58, pp. 235–6] [29, Secs. 253–254] But this technique is explicitly worked out in Maclaurin's *Treatise of Fluxions*. Indeed, it appears twice: once in geometric dress in Book I, Chapter IX, and then more algebraically in Book II. [63, pp. 694–696] Euler, in the version he gave in his 1755 textbook, [20] does not refer to Maclaurin on this point, but then he makes few references in that book at all. Still we might suspect, especially knowing that Stirling told Euler in a letter of 16 April 1738 [91] that Maclaurin had some interesting results on series, that Euler would have been particularly interested in looking at Maclaurin's applications of the Taylor series. Certainly Lacroix's praise for Maclaurin's work on series must have taken this set of results into account. [52, p. xxvii] Even more important, Lagrange, in unpublished lectures on the calculus from Turin in the 1750's, after giving a very elementary treatment of maxima and minima, referred to volume II of Maclaurin's *Treatise of Fluxions* as the chief source for more information on the subject. [7, p. 154] Since Lagrange did not mention Euler in this connection at all, Lagrange could well not even have seen the *Institutiones calculi differentialis* of 1755 when he made this reference. This Taylor-series approach to maxima and minima (with the Lagrange remainder supplied for the Taylor series) plays a major role in the work

of Lagrange, and later in the work of Cauchy. It is because Maclaurin thought of maxima and minima, and of convexity and concavity, in Archimedean geometrical terms that he was led to look at the relevant inequalities, just as the geometry of Archimedes helped Maclaurin formulate some of the inequalities he used to prove his special case of the Fundamental Theorem of Calculus.

b. Ellipsoids We now turn to work in applied mathematics that constitutes one of Maclaurin's great claims to fame: the gravitational attraction of ellipsoids and the related problem of the shape of the earth. Maclaurin is still often regarded as the creator of the subject of attraction of ellipsoids. [85, pp. 175, 374] In the eighteenth century, the topic attracted serious work from d'Alembert, A.-C. Clairaut, Euler, Laplace, Lagrange, Legendre, Poisson, and Gauss. In the twentieth century, Subramanyan Chandrasekhar (later Nobel laureate in physics) devoted an entire chapter of his classic *Ellipsoidal Figures of Equilibrium* to the study of Maclaurin spheroids (figures that arise when homogeneous bodies rotate with uniform angular velocity), the conditions of stability of these spheroids and their harmonic modes of oscillation, and their status as limiting cases of more general figures of equilibrium. Such spheroids are part of the modern study of classical dynamics in the work of scientists like Chandrasekhar, Laurence Rossner, Carl Rosenkilde, and Norman Lebovitz. [15, pp. 77–100] Already in 1740 Maclaurin had given a "rigorously exact, geometrical theory" of homogeneous ellipsoids subject to inverse-square gravitational forces, and had shown that an oblate spheroid is a possible figure of equilibrium under Newtonian mutual gravitation, a result with obvious relevance for the shape of the earth. [39, p. 172] [86 p. xix] [85, p. 374]

Of particular importance was Maclaurin's decisive influence on Clairaut. Maclaurin and Clairaut corresponded extensively, and Clairaut's seminal 1743 book *La Figure de la Terre* [18] frequently, explicitly, and substantively cites his debts to Maclaurin's work. [39, pp. 590–597] A key result, that the attractions of two confocal ellipsoids at a point external to both are proportional to their masses and are in the same direction, was attributed to Maclaurin by d'Alembert, an attribution repeated by Laplace, Lagrange, and Legendre, then by Gauss, who went back to Maclaurin's original paper, and finally by Lord Kelvin, who called it "Maclaurin's splendid theorem." [15, p. 38] [85, pp. 145, 409] Lagrange

began his own memoir on the attraction of ellipsoids by praising Maclaurin's treatment in the prize paper of 1740 as a masterwork of geometry, comparing the beauty and ingenuity of Maclaurin's work with that of Archimedes, [57, p. 619] though Lagrange, typically, then treated the problem analytically. Maclaurin's eighteenth- and nineteenth-century successors also credit him with some of the key methods used in studying the equilibrium of fluids, such as the method of balancing columns. [39, p. 597] Maclaurin's work on the attraction of ellipsoids shows how his geometric insights fruitfully influenced a subject that later became an analytic one.

c. The Euler-Maclaurin formula The Euler-Maclaurin formula expresses the value of definite integrals by means of infinite series whose coefficients involve what are now called the Bernoulli numbers. The formula shows how to use integrals to find the partial sums of series. Maclaurin's version, in modern notation, is:

$$\sum_{h=0}^{\infty} F(a+h) = \int_0^a F(x) dx + \frac{1}{2}F(a) + \frac{1}{2}F'(a) - \frac{1}{720}F'''(a) + \frac{1}{30240}F^{(v)}(a) - \dots$$

[35, pp. 84–86] James Stirling in 1738, congratulating Euler on his publication of that formula, told Euler that Maclaurin had already made it public in the first part of the *Treatise of Fluxions*, which was printed and circulating in Great Britain in 1737. [47, p. 88n] [91, p. 178] (On this early publication, see also [63, p. iii, p. 691n].) P. L. Griffiths has argued that this simultaneous discovery rests on De Moivre's work on summing reciprocals, which also involves the so-called Bernoulli numbers. [40] [41, pp. 16–17] [25, p. 19] In any case, Euler and Maclaurin derived the Euler-Maclaurin formula in essentially the same way, from a similar geometric diagram and then by integrating various Taylor series and performing appropriate substitutions to find the coefficients. [31] [32] [33] Maclaurin's approach is no more Archimedean or geometric than Euler's; they are similar and independent. [63, pp. 289–293, 672–675] [35, pp. 84–93] [67] In subsequent work, Euler went on to extend and apply the formula further to many other series, especially in his *Introductio in Analysisin Infinitorum* of 1748 and *Institutiones Calculi Differentialis* of 1755. [35, p. 127] But Maclaurin, like Euler, had applied the formula to solve many problems. [63, pp. 676–693]

For instance, Maclaurin used it to sum powers of arithmetic progressions and to derive Stirling's formula for factorials. He also derived what is now called the Newton-Cotes numerical integration formula, and obtained what is now called Simpson's rule as a special case. It is possible that his work helped stimulate Euler's later, fuller investigations of these important ideas.

In 1772, Lagrange generalized the Euler-Maclaurin formula, which he obtained as a consequence of his new calculus of operators. [53] [35, pp. 169, 261] In 1834, Jacobi provided the formula with its remainder term, [46, p. 263, 265] in the same paper in which he first introduced what are now called the Bernoulli polynomials. Jacobi, who called the result simply the Maclaurin summation formula, cited it directly from the *Treatise of Fluxions*. [46, p. 263] Later, Karl Pearson used the formula as an important tool in his statistical work, especially in analyzing frequency curves. [72, pp. 217, 262]

The Euler-Maclaurin formula, then, is an important result in the mainstream of mathematics, with many applications, for which Maclaurin, both in the eighteenth century and later on, has rightly shared the credit.

d. Elliptic integrals Some integrals (Maclaurin used the Newtonian term “fluents”), are algebraic functions, Maclaurin observed. Others are not, but some of these can be reduced to finding circular arcs, others to finding logarithms. By analogy, Maclaurin suggested, perhaps a large class of integrals could be studied by being reduced to finding the length of an elliptical or hyperbolic arc. [63, p. 652] By means of clever geometric transformations, Maclaurin was able to reduce the integral that represented the length of a hyperbolic arc to a ‘nice’ form. Then, by algebraic manipulation, he could reduce some previously intractable integrals to that same form. His work was translated into analysis by d'Alembert and then generalized by Euler. [13, p. 846] [23] [27, p. 526] [28, p. 258] In 1764, Euler found a much more elegant, general, and analytic version of this approach, and worked out many more examples, but cited the work of Maclaurin and d'Alembert as the source of his investigation. A.-M. Legendre, the key figure in the eighteenth-century history of elliptic integrals, credited Euler with seeing that, by the aid of a good notation, arcs of ellipses and other transcendental curves could be as generally used in integration as circular and logarithmic arcs. [45, p. 139] Legendre was, of course, right that “elliptic integrals” en-

compass a wide range of examples; this was exactly Maclaurin's point. Thus, although his successors accomplished more, Maclaurin helped initiate a very important investigation and was the first to appreciate its generality. Maclaurin's geometric insight, applied to a problem in analysis, again brought him to a discovery.

7 Other examples of Maclaurin's mathematical influence

The foregoing examples provide evidence of direct influence of the *Treatise of Fluxions* on Continental mathematics. There is much more. For instance, Lacroix, in his treatment of integrals by the method of partial fractions, called it "the method of Maclaurin, followed by Euler." [52, Vol. II, p. 10] [63, pp. 634–644] Of interest too is Maclaurin's clear understanding of the use of limits in founding the calculus, especially in the light of his likely influence on d'Alembert's treatment of the foundations of the calculus by means of limits in the *Encyclopédie*, which in turn influenced the subsequent use of limits by L'Huilier, Lacroix, and Cauchy, [38, ch. 3] (and on Lagrange's acceptance of the limit approach in his early work in the 1750's). [7] Although the largest part of Maclaurin's reply to Berkeley was the extensive proof of results in calculus using Greek methods, he was willing to explain important concepts using limits also. In particular, Maclaurin wrote, "As the tangent of an arch [arc] is the right line that limits the position of all the secants that can pass through the point of contact . . . though strictly speaking it be no secant; so a ratio may limit the variable ratios of the increments, though it cannot be said to be the ratio of any real increments." [63, p. 423] Maclaurin's statement answers Berkeley's chief objection—that the increment in a function's value is first treated as non-zero, then as zero, when one calculates the limit of the ratio of increments or finds the tangent to a curve. Maclaurin's statement is in the tradition of Newton's *Principia* (Book I, Scholium to Lemma XI), but is in a form much closer to the later work of d'Alembert on secants and tangents. [20] Maclaurin pointed out that most of the propositions of the calculus that he could prove by means of geometry "may be *briefly* demonstrated by this method [of limits]." [63, p. 87, my italics]

In addition, Maclaurin had considerable influence in Britain, on mathematicians like John Landen (whose work on series was praised by Lagrange),

Robert Woodhouse (who sparked the new British interest in Continental work about 1800), and on Edward Waring and Thomas Simpson, whose names are attached to results well known today. [42] Going beyond the calculus, Maclaurin's purely geometric treatises were read and used by French geometers of the stature of Chasles and Poncelet. [90, p. 145] Thus, though Maclaurin may not have been the towering figure Euler was, he was clearly a significant and respected mathematician, and the *Treatise of Fluxions* was far more than an unread tome whose weight served solely to crush Bishop Berkeley.

8 Why a *Treatise of Fluxions*?

The *Treatise of Fluxions* was not really intended as a reply to Berkeley. Maclaurin could have refuted Berkeley with a pamphlet. It was not a student handbook either; this work is far from elementary. Nor was it merely written to glory in Greek geometry. Maclaurin wrote several works on geometry per se. But he was no antiquarian. Instead, the *Treatise of Fluxions* was the major outlet for Maclaurin's solution of significant research problems in the field we now call analysis. Geometry, as the examples I gave illustrate, was for Maclaurin a source of motivation, of insight, and of problem-solving power, as well as being his model of rigor.

For Maclaurin, rigor was not an end in itself, or a goal pursued for purely philosophical reasons. It was motivated by his research goals in analysis. For instance, Maclaurin developed his theory of maxima, minima, points of inflection, convexity and concavity, orders of contact, etc., because he wanted to study curves of all types, including those that cross over themselves, loop around and are tangent to themselves, and so on. He needed a sophisticated theory to characterize the special points of such curves. Again, in problems as different as studying the attraction of ellipsoids and evaluating integrals approximately, he needed to use infinite series and know how close he was to their sum. Thus, rigor, to Maclaurin, was not merely a tool to defend Newton's calculus against Berkeley—though it was that—nor just a response to the needs of a professor to present his students a finished subject—though it may have been that as well. In many examples, Maclaurin's rigor serves the needs of his research.

Moreover, the *Treatise of Fluxions* contains a wealth of applications of fluxions, from standard physical problems such as curves of quickest de-

scent to mathematical problems like the summation of power series—in the context of which, incidentally, Maclaurin gave what may be the earliest clear definition of the sum of an infinite series: “There are progressions of fractions which may be continued at pleasure, and yet the sum of the terms be always less than a certain finite number. If the difference betwixt their sum and this number decrease in such a manner, that by continuing the progression it may become less than any fraction how small soever that can be assigned, this number is the *limit of the sum of the progression*, and is what is understood by the value of the progression when it is supposed to be continued indefinitely.” [63, p. 289] Thus, though eighteenth-century Continental mathematicians did not care passionately about foundations, [38, pp. 18–24] they could still appreciate the *Treatise of Fluxions* because they could mine it for results and techniques.

9 Why the traditional view?

If the reader is convinced by now that the traditional view is wrong, that Maclaurin's *Treatise* did not mark the end of the Newtonian tradition, and that not all of modern analysis stems solely from the work of Leibniz and his school, the question arises, how did that traditional view come to be, and why it has been so persistent?

Perhaps the traditional view could be explained as follows. Consider the approach to mathematics associated with Descartes: symbolic power, not debates over foundations; problem-solving power, not axioms or long proofs. The Cartesian approach to mathematics is clearly reflected in the work and in the rhetoric of Leibniz, Johann Bernoulli, Euler, Lagrange—especially in the historical prefaces to his influential works—and even Cauchy. These men, the giants of their time, are linked in a continuous chain of teachers, close colleagues, and students. Some topics, like partial differential equations and the calculus of variations, were developed mostly on the Continent. Moreover, the Newton-Leibniz controversy helped drive English and Continental mathematicians apart. Thus the Continental tradition can be viewed as self-contained, and the outsider sees no need for eighteenth-century Continental mathematicians to struggle through 750 pages of a *Treatise of Fluxions*, which is at best in the Newtonian notation and at worst in the language of Greek geometry. Lagrange's well-known boast that his *Analytical*

Mechanics [55] had (and needed) no diagrams, thus opposing analysis to geometry at the latter's expense, reinforced these tendencies and enshrined them in historical discourse. But the explanation we have just given does not suffice to explain the strength, and persistence into the twentieth century, of the standard interpretation. The traditional view of Maclaurin's lack of importance has been reinforced by some other historiographical tendencies that deserve our critical attention.

The traditional picture of Maclaurin's *Treatise of Fluxions* radically separates his work on foundations, which it regards as geometric, sterile, and antiquarian, from his important individual results, which often are mentioned in histories of mathematics but are treated in isolation from the purpose of the *Treatise*, in isolation from one another, and in isolation from Maclaurin's overall approach to mathematics. Strangely, both externalist and internalist historians, each for different reasons, have reinforced this picture.

For instance, in the English-speaking world, viewing the *Treatise* as only about Maclaurin's foundation for the calculus, and thus as a dead end, has been perpetuated by the “decline of science in England” school of the history of eighteenth-century science, stemming from such early nineteenth-century figures as John Playfair, and, especially, Charles Babbage. [77] [2] [4] Babbage felt strongly about this because he was a founder of the Cambridge Analytical Society, which fought to introduce Continental analysis into Cambridge in the early nineteenth century. This group had an incentive to exaggerate the superiority of Continental mathematics and downgrade the British, as is exemplified by their oft-quoted remark that the principles of “pure d-ism” should replace what they called the “dot-age” of the University. [5, ch. 7] [10, p. 274] The pun, playing on the Leibnizian and Newtonian notation in calculus, may be found in [2, p. 26]. These views continued to be used in the attempt by Babbage and others to reform the Royal Society and to increase public support for British science.

It is both amusing and symptomatic of the misunderstanding of Maclaurin's influence that Lacroix's one-volume treatise on the calculus of 1802, [50] translated into English by the Cambridge Analytical Society with added notes on the method of series of Lagrange, [51] was treated by them, and has been considered since, as a purely “Continental” work. But Lacroix's short treatise was based on the concept of limit, which was Newtonian, elaborated by

Maclaurin, adapted by d'Alembert and L'Huilier, and finally systematized by Lacroix. [38, pp. 81–86] Moreover, the translators' notes by Babbage, Herschel, and Peacock supplement the text by studying functions by their Taylor series, thus using the approach that Lacroix himself, in his multi-volume treatise of 1810, had attributed to Maclaurin. This is, of course, not to deny the overwhelming importance of the contributions of Euler and Lagrange, both to the mathematics taught by the Analytical Society and to that included by Lacroix in his 1802 book, nor to deny the Analytical Society's emphasis on a more abstract and formal concept of function. But all the same, Babbage, Herschel, and Peacock were teaching some of Maclaurin's ideas without realizing this.

In any case, the views expressed by Babbage and others have strongly influenced Cambridge-oriented writers like W. W. Rouse Ball, who said that the history of eighteenth-century English mathematics "leads nowhere." [5, p. 98] H. W. Turnbull, though he wrote sympathetically about Maclaurin's mathematics on one occasion, [88] blamed Maclaurin on another occasion for the decline: "When Maclaurin produced a great geometrical work on fluxions, the scale was so heavily loaded that it diverted England from Continental habits of thought. During the remainder of the century, British mathematics were relatively undistinguished." [89, p. 115]

Historians of Scottish thought, working from their central concerns, have also unintentionally contributed to the standard picture. George Elder Davie, arguing from social context to a judgment of Maclaurin's mathematics, held that the Scots, unlike the English, had an anti-specialist intellectual tradition, based in philosophy, and emphasizing "cultural and liberal values." Wishing to place Maclaurin in this context, Davie stressed what he called Maclaurin's "mathematical Hellenism," [24, p. 112] and was thus led to circumscribe the achievement of the *Treatise of Fluxions* as having based the calculus "on the Euclidean foundations provided by [Robert] Simson," [24, p. 111] who had made the study of the writings of the classical Greek geometers the "national norm" in Scotland. The "Maclaurin is a geometer" interpretation among Scottish historians has been further reinforced by a debate in 1838 over who would fill the Edinburgh chair in mathematics. Phillip Kelland, a candidate from Cambridge, was seen as the champion of Continental analysis, while the partisans of Duncan Gregory argued for a more geometrical approach. Wishing to enlist the entire Scottish geometric tradition on the side of Gre-

gory, Sir William Hamilton wrote, "The great Scottish mathematicians, . . . even Maclaurin, were decidedly averse from the application of the mechanical procedures of algebra." [24, p. 155] Though Kelland eventually won the chair, the dispute helped spread the view that Maclaurin had been hostile to analysis. More recently, Richard Olson has characterized Scottish mathematics after Maclaurin as having been conditioned by Scottish common-sense philosophy to be geometric in the extreme. [70, pp. 4, 15] [71, p. 29] But in emphasizing Maclaurin's influence on this development, Olson, like Davie, has overstated the degree to which Maclaurin's approach was geometric.

By contrast, consider internalist historians. The treatment of Maclaurin's results as isolated reflects what Herbert Butterfield called the Whig approach to history, viewing the development of eighteenth-century mathematics as a linear progression toward what we value today, the collection of results and techniques which make up classical analysis. Thus, mathematicians writing about the history of this period, from Moritz Cantor in the nineteenth century to Hermann Goldstine and Morris Kline in the twentieth, tell us what Maclaurin did with specific results, some named after him, for which they have mined the *Treatise of Fluxions*. [13, pp. 655–663] [35, pp. 126ff, 167–168] [49, pp. 522–523, 452, 442] They either neglect the apparently fruitless work on foundations, or, viewing it as geometric, see it as a step backward. It is of course true that many Continental mathematicians used Maclaurin's results without accepting the geometrical and Newtonian insights that Maclaurin used to produce them. But without those points of view, Maclaurin would not have produced those results.

Both externalist and internalist historians, then, have treated Maclaurin's work in the same way: as a throwback to the Greeks, with a few good results that happen to be in there somewhat like currants in a scone. Further, the fact that Maclaurin's book, especially its first hundred pages, is very hard to read, especially for readers schooled in modern analysis, has encouraged historians who focus on foundations to read only the introductory parts. The fact that there is so much material has encouraged those interested in results to look only at the sections of interest to them. And the fact that the first volume is so overwhelmingly geometric serves to reinforce the traditional picture once again whenever anybody opens the *Treatise*. The recent Ph.D. dissertation by Erik Sageng [78] is the first example of a modern

scholarly study of Maclaurin's *Treatise* in any depth. The standard picture has not yet been seriously challenged in print.

10 Some final reflections

Maclaurin's work had Continental influence, but with an important exception—his geometric foundation for the calculus. Mastering this is a major effort, and I know of no evidence that any eighteenth-century Continental mathematician actually did so. Lagrange perhaps came the closest. In the introduction to his *Théorie des fonctions analytiques*, Lagrange could say only, Maclaurin did a good job basing calculus on Greek geometry, so it can be done, but it is very hard. [58, p. 17] In an unpublished draft of this introduction, Lagrange said more pointedly: "I appeal to the evidence of all those with the courage to read the learned treatise of Maclaurin and with enough knowledge to understand it: have they, finally, had their doubts cleared up and their spirit satisfied?" [73, p. 30]

Something else may have blunted people's views of the mathematical quality of Maclaurin's *Treatise*. The way the book is constructed partly reflects the Scottish intellectual milieu. The Enlightenment in Britain, compared with that on the Continent, was marked less by violent contrast and breaks with the past than by a spirit of bridging and evolution. [75, pp. 7–8, 15] Similarly, Scottish reformers operated less by revolution than by the refurbishment of existing institutions. [16, p. 8] These trends are consistent with the two-fold character of the *Treatise of Fluxions*: a synthesis of the old and the new, of geometry and algebra, of foundations and of new results, a refurbishment of Newtonian fluxions to deal with more modern problems. This contrasts with the explicitly revolutionary philosophy of mathematics of Descartes and Leibniz, and thus with the spirit of the mathématicien of the eighteenth century on the Continent.

Of course Scotland was not unmarked by the conflicts of the century. During the Jacobite rebellion in 1745, Maclaurin took a major role in fortifying Edinburgh against the forces of Bonnie Prince Charlie. When the city was surrendered to the rebels, Maclaurin fled to York. Before his return, he became ill, and apparently never really recovered. He briefly resumed teaching, but died in 1746 at the relatively young age of forty-eight. Nonetheless, the Newtonian tradition in the calculus was not a dead

end. Maclaurin in his lifetime, and his *Treatise of Fluxions* throughout the century, transmitted an expanded and improved Newtonian calculus to Continental analysts. And Maclaurin's geometric insight helped him advance analytic subjects.

We conclude with the words of an eighteenth-century Continental mathematician whose achievements owe much to Maclaurin's work. [39, pp. 172, 412–425, 590–597] The quotation [66, p. 350] illustrates Maclaurin's role in transmitting the Newtonian tradition to the Continent, the respect in which he was held, and the eighteenth-century social context essential to understanding the fate of his work. In 1741, Alexis-Claude Clairaut wrote to Colin Maclaurin, "If Edinburgh is, as you say, one of the farthest corners of the world, you are bringing it closer by the number of beautiful discoveries you have made."

References

1. Arnold, Matthew, The Literary Importance of Academies, in Matthew Arnold, *Essays in Criticism*, Macmillan, London, 1865, 42–79. Cited in [61].
2. Babbage, Charles, *Passages from the Life of a Philosopher*, Longman, London, 1864.
3. Babbage, Charles, Preface to *Memoirs of the Analytical Society* (Cambridge: J. Smith, 1813). Attributed to Babbage by Anthony Hyman, *Charles Babbage: Pioneer of the Computer*, Princeton University Press, Princeton, 1982.
4. —, *Reflections on the Decline of Science in England, and Some of its Causes*, B. Fellowes, London, 1830.
5. Ball, W. W. Rouse, *A History of the Study of Mathematics at Cambridge*, Cambridge University Press, Cambridge, 1889.
6. Berkeley, George, *The Analyst, or a Discourse Addressed to an Infidel Mathematician*, in A. A. Luce and T. R. Jessop, eds., *The Works of George Berkeley*, vol. 4, T. Nelson, London, 1951, 65–102.
7. Borgato, Maria Teresa, and Luigi Pepe, Lagrange a Torino (1750–1759) e le sue lezioni inedite nelle R. Scuole di Artiglieria, *Bollettino di Storia delle Scienze Matematiche* 1987, 7: 3–180.
8. Bourbaki, Nicolas, *Elements d'Histoire des Mathématiques*, Paris: Hermann, Paris, 1960.
9. Boyer, Carl, *The History of the Calculus and Its Conceptual Development*, Dover, New York, 1959.
10. Cajori, Florian, *A History of the Conceptions of Limits and Fluxions in Great Britain from Newton to Woodhouse*, Open Court, Chicago and London, 1919.

11. —, *A History of Mathematics*, 2d, ed., Macmillan, New York, 1922.
12. Cantor, G. N., Anti-Newton, in J. Fauvel et al, eds., *Let Newton Be!*, Oxford University Press, Oxford, 1988, pp. 203–222.
13. Cantor, Moritz, *Vorlesungen über Geschichte der Mathematik*, vol. 3, Teubner, Leipzig, 1898.
14. Cauchy, A.-L., *Résumé des leçons données à l'école royale polytechnique sur le calcul infinitésimal*, in *Oeuvres complètes*, Ser. 2, vol. 4, Gauthier-Villars, Paris, 1899.
15. Chandrasekhar, S., *Ellipsoidal Figures of Equilibrium*, Yale, New Haven, 1969.
16. Chitnis, Anand, *The Scottish Enlightenment: A Social History*, Croom Helm, London, 1976.
17. Christie, John R. R., The Origins and Development of the Scottish Scientific Community, 1680–1760, *History of Science* 12 (1974), 122–141.
18. Clairaut, A.-C., *Théorie de la Figure de la Terre*, Durand, Paris, 1743.
19. d'Alembert, Jean, Différentiel, in [21].
20. d'Alembert, Jean, and de la Chapelle, Limite, in [21].
21. d'Alembert, Jean, et al, eds., *Dictionnaire Encyclopédique des Mathématiques*, Hotel de Thou, Paris, 1789, which collects the mathematical articles from the Diderot-d'Alembert *Encyclopédie*.
22. d'Alembert, Jean, *Traité de Dynamique*, David Ian, Paris, 1743.
23. —, Recherches sur le calcul intégral, *Histoire de l'Académie de Berlin* (1746), 182–224.
24. Davie, George Elder, *The Democratic Intellect: Scotland and her Universities in the Nineteenth Century*, The University Press, Edinburgh, 1966.
25. De Moivre, Abraham, *Miscellanea analytica*, Tonson and Watts, London, 1730.
26. Dijksterhuis, E. J., *Archimedes*, 2nd ed., tr. C. Dikshoorn, Princeton University Press, Princeton, 1987.
27. Enneper, Alfred, *Elliptische Functionen: Theorie und Geschichte*, 2nd ed., Nebert, Halle, 1896.
28. Euler, Leonhard, De reductione formularum integralium ad rectificationem ellipsis ac hyperbolae, *Nov. Comm. Petrop.* 10 (1764), 30–50, in L. Euler, *Opera Omnia*, Teubner, Leipzig, Berlin, Zurich, 1911–, Ser. I, vol. 20, 256–301.
29. —, *Institutiones Calculi Differentialis*, 1755, sections 253–255. In *Opera*, Ser. I, Vol XI.
30. —, *Introductio in Analysin Infinitorum*, Lausanne, 1748, in *Opera*, Ser. I, vols. 8–9.
31. —, Inventio summae cuiusque seriei ex data termino generali, *Comm. Petrop.* 8 (1741), 9–22; in *Opera*, Ser. I, vol. 14, 108–123.
32. —, Methodus generalis summandi progressionis, *Comentarii Acad. Imper. Petrop.* 6 (1738), 68–97; in *Opera*, Ser. II, vol. 22.
33. Euler, Leonhard, Methodus universalis serierum convergentium summas quam proxime inveniendi, *Comm. Petrop.* 8 (1741), 3–9, in *Opera*, Ser. I, vol. 14, 101–107.
34. Eyles, V. A., Hutton, *Dictionary of Scientific Biography*, vol. 6, 577–589.
35. Goldstine, Herman, *A History of Numerical Analysis from the 16th through the 19th Century*, Springer-Verlag, New York, Heidelberg, Berlin, 1977.
36. Grabiner, Judith V., A Mathematician Among the Molasses Barrels: MacLaurin's Unpublished Memoir on Volumes, *Proceedings of the Edinburgh Mathematical Society* 39 (1996), 193–240.
37. —, *The Calculus as Algebra: J.-L. Lagrange, 1736–1813*, Garland Publishing, Boston, 1990.
38. —, *The Origins of Cauchy's Rigorous Calculus*, MIT Press, Cambridge, Mass., 1981.
39. Greenberg, John L., *The Problem of the Earth's Shape from Newton to Clairaut*, Cambridge University Press, Cambridge, 1995.
40. Griffiths, P. L., Private communication.
41. —, The British Influence on Euler's Early Mathematical Discoveries, preprint.
42. Guicciardini, Niccolo, *The Development of Newtonian Calculus in Britain, 1700–1800*, Cambridge University Press, Cambridge, 1989.
43. Hall, A. Rupert, *Philosophers at War: The Quarrel between Newton and Leibniz*, Cambridge University Press, Cambridge, 1980.
44. Hankins, Thomas, *Jean d'Alembert: Science and the Enlightenment*, Clarendon Press, Oxford, 1970.
45. Itard, Jean, Legendre, *Dictionary of Scientific Biography*, vol. 8, 135–143.
46. Jacobi, C. G. J., De usu legitimo formulae summatoriae Maclauriniana, *Journ. f. reine u. angew. Math.* 18 (1834), 263–272. Also in *Gesammelte Werke*, vol. 6, 1891, pp. 64–75.
47. Juškevič, A. P., and R. Taton, eds., *Leonhard Euleri commercium Epistolicum*, Birkhauser, Basel, 1980. In Leonhard Euler, *Opera*, Ser. 4, vol. 5.
48. Juškevič, A. P., and Winter, E., eds., *Leonhard Euler und Christian Goldbach: Briefwechsel, 1729–1764*, Akademie-Verlag, Berlin, 1965.
49. Kline, Morris, *Mathematical Thought from Ancient to Modern Times*, Oxford, New York, 1972.
50. Lacroix, S. F., *Traité Élémentaire de Calcul Différentiel et de Calcul Intégral*, Duprat, Paris, 1802. Translated as [51].

51. —, *An Elementary Treatise on the Differential and Integral Calculus*, translated, with an Appendix and Notes, by C. Babbage, J. F. W. Herschel, and G. Peacock, J. Deighton and Sons, Cambridge, 1816.
52. —, *Traité du calcul différentiel et du calcul intégral*, 3 vols., 2nd ed., Courcier, Paris, 1810–1819, Vol. I, 1810.
53. Lagrange, J.-L., Sur une nouvelle espèce de calcul relatif à la différentiation et à l'intégration des quantités variables, *Nouvelles Mémoires de l'Académie . . . de Berlin*, 1772, 185–221; in *Oeuvres*, vol. 3, 439–476.
54. —, *Leçons sur le Calcul des Fonctions*, new ed. Courcier, Paris, 1806; in *Oeuvres*, vol. 10.
55. —, *Mécanique analytique*, 2d. ed., 2 vols., Courcier, Paris, 1811–1815; in *Oeuvres*, vols. 11–12.
56. —, Note sur la métaphysique du calcul infinitésimal, *Miscellanea Taurinensia* 2 (1760–61), 17–18; in *Oeuvres*, vol. 7, 597–599.
57. —, Sur l'attraction des sphéroïdes elliptiques, *Mémoires de l'académie de Berlin*, 1773, 121–148. Reprinted in *Oeuvres de Lagrange*, vol. III, 619ff.
58. —, *Théorie des fonctions analytiques*, Imprimerie de la République, Paris, An V [1797]; compare the second edition, Courcier, Paris, 1813, reprinted in *Oeuvres de Lagrange*, pub. M. J.-A. Serret, 14 volumes, Gauthier-Villars, Paris, 1867–1892, reprinted again, Georg Olms Verlag, Hildesheim and New York, 1973, vol. 9.
59. Legendre, Adrien-Marie, Mémoires sur les intégrations par arcs d'ellipse et sur la comparaison de ces arcs, *Mémoires de l'Académie des Sciences*, 1786, 616, 644–673.
60. —, *Traité des Fonctions Elliptiques et des Intégrales Eulériennes, avec des Tables pour en faciliter le Calcul Numérique*, 3 vols., Paris, 1825–1828.
61. Loria, Gino, The Achievements of Great Britain in the Realm of Mathematics, *Mathematical Gazette* 8 (1915), 12–19.
62. Maclaurin, Colin, *Traité de Fluxions*, Traduit de l'anglois par le R. P. Pèzénas, 2 vols., Jombert, Paris, 1749.
63. —, *A Treatise of Fluxions in Two Books*, Ruddimans, Edinburgh, 1742.
64. Mahoney, Michael, Review of [42], *Science* 250 (1990), 144.
65. McElroy, Davis, *Scotland's Age of Improvement*, Washington State University Press, Pullman, 1969.
66. Mills, Stella, *The Collected Letters of Colin Maclaurin*, Shiva Publishing, Nantwich, 1982.
67. —, The Independent Derivations by Leonhard Euler and Colin Maclaurin of the Euler-Maclaurin Summation Formula, *Archive for History of Exact Science* 33 (1985), 1–13.
68. Murdoch, Patrick, An Account of the Life and Writings of the Author, in Colin Maclaurin, *Account of Sir Isaac Newton's Philosophical Discoveries*, For the Author's Children, London, 1748, i–xx; reprinted, Johnson Reprint Corp., New York, 1968.
69. Newton, Isaac, *Of Analysis by Equations of an Infinite Number of Terms*, J. Stewart, London, 1745, in D. T. Whiteside, ed., *Mathematical Works of Isaac Newton*, vol. I, Johnson Reprint, New York and London, 1964, 3–25.
70. Olson, Richard, *Scottish Philosophy and British Physics, 1750–1880*, Princeton University Press, Princeton, 1975.
71. —, Scottish Philosophy and Mathematics, 1750–1830, *Journal of the History of Ideas* 32 (1971), 29–44.
72. Pearson, Karl, *The History of Statistics in the Seventeenth and Eighteenth Centuries* [written 1921–1933]; ed. E. S. Pearson, Charles Griffin & Co., London and High Wycombe, 1976.
73. Pepe, Luigi, Tre 'prime edizioni' ed un' introduzione inedita della Fonctions analytiques di Lagrange, *Boll. Stor. Sci. Mat.* 6 (1986), 17–44.
74. Phillipson, Nicholas, The Scottish Enlightenment, in [76], pp. 19–40.
75. Porter, Roy, The Enlightenment in England, in [76], pp. 1–18.
76. Porter, Roy, and Mikulas Teich, eds., *The Enlightenment in National Context*, Cambridge University Press, Cambridge, 1981.
77. Playfair, John, *Traité de Méchanique Celeste*, *Edinburgh Review* 22 (1808), 249–84.
78. Sageng, Erik Lars, *Colin Maclaurin and the Foundations of the Method of Fluxions*, unpublished Ph. D. Dissertation, Princeton University, 1989.
79. Scott, J. F., Maclaurin, *Dictionary of Scientific Biography*, vol. 8, 609–612.
80. Shapin, Stephen, and Arnold Thackray, Prosopography as a Research Tool in History of Science: The British Scientific Community, 1700–1900, *History of Science* 12 (1974), 95–121.
81. Stewart, M. A., ed., *Studies in the Philosophy of the Scottish Enlightenment*, Clarendon Press, Oxford, 1990.
82. Struik, D. J., *A Source Book in Mathematics, 1200–1800*, Harvard University Press, Cambridge, Mass., 1969.

83. Taton, Juliette, Pezenas, *Dictionary of Scientific Biography*, vol. 10, Scribners, New York, 1974, pp. 571–572.
84. Taton, René, ed., *Enseignement et Diffusion des Sciences en France au XVIII^e Siècle*, Hermann, Paris, 1964.
85. Todhunter, Isaac, *A History of the Mathematical Theories of Attraction and the Figure of the Earth, from the Time of Newton to that of Laplace*, Macmillan, London, 1873.
86. Truesdell, C., Rational Fluid Mechanics, 1687–1765, introduction to Euler, *Opera*, Ser. 2, vol. 12.
87. ———, The Rational Mechanics of Flexible or Elastic Bodies, 1638–1788, in Euler, *Opera*, Ser. 2, vol. 11.
88. Turnbull, H. W., *Bicentenary of the Death of Colin Maclaurin (1698–1746)*, The University Press, Aberdeen, 1951.
89. ———, *The Great Mathematicians*, Methuen, London, 1929.
90. Tweedie, Charles, A Study of the Life and Writings of Colin Maclaurin, *Mathematical Gazette* 8 (1915), 132–151.
91. ———, *James Stirling*, Clarendon Press, Oxford, 1922.

Discussion of Fluxions: from Berkeley to Woodhouse

FLORIAN CAJORI

American Mathematical Monthly 24 (1917), 145–154

The first direct statement of Newton's method and notation of fluxions was printed in 1693 in Wallis's *Algebra*. Here and in the *Principia* of 1687 Newton made use of infinitely small quantities, but in his *Quadrature of Curves* of 1704 he declared that "in the method of fluxions there is no necessity of introducing figures infinitely small." No other publication of Newton, printed either before 1704 or after, equalled the *Quadrature of Curves* in mathematical rigor. Here Newton reached his high water mark of rigidity in the exposition of fluxions. By a fluxion, Newton always meant a finite velocity. With one exception, all British writers on the new calculus before the appearance of Berkeley's *Analyst* in 1734 used the Newtonian notation consisting of dots or "prick'd letters," and also Newton's word "fluxion." But strange to say, most of these writers did not use Newton's concepts. They applied the term "fluxion" to the infinitely small quantities of Leibniz—thus using a home label on goods of foreign manufacture. Of sixteen or more writers in Great Britain during the period of 1693–1734, nine or more call a fluxion an infinitely small quantity; three writers do not define their terms, while only four follow Newton's exposition of 1687 or 1693, involving fluxions as finite velocities and "moments" as infinitely small quantities, or else follow Newton's exposition of 1704, involving fluxions as finite velocities and avoiding infinitely small quantities almost entirely. The nine or more who used fluxions in the sense of infinitely small quantities had no hesitation in dropping quantities from an equation when they were very small in comparison with the other quantities. Altogether these writings contained a medley of philosophical doctrine which presented

a great opportunity for destructive criticism on the part of such a close reasoner and skilful debater as Bishop Berkeley. Before this no mathematical subject, except Zeno's paradoxes on motion, had ever offered itself as a topic for picturesque dialectics; before Berkeley only once was such expert and splendid dialectical energy brought to bear on a fundamental topic in mathematics.

Berkeley's *Analyst* marks a turning point in the history of mathematical thought in Great Britain. His criticisms were not openly accepted by mathematicians of his day; nevertheless such effort was put forth to avoid his objections that in eight years the logical exposition of fluxions was immensely improved.

In the library of Trinity College, Cambridge, there is a marble bust of James Jurin, a noted physician, at one time a student at Trinity. He undertook a defence of Newton and the calculus. Under the pseudonym of "Philalethes Cantabrigiensis," Jurin wrote two long replies to Berkeley, full of noisy rhetoric and giving little that was truly substantial. Berkeley found a second antagonist in John Walton, a professor of mathematics in Dublin, who had a good intuitive grasp of fluxions, but lacked deep philosophical insight and showed himself inexperienced in the conduct of controversies. Walton wrote two replies to Berkeley and an augmentation of his second reply. Altogether this discussion involved eight articles, three by Berkeley, two by Jurin and three by Walton.

Berkeley made some mistakes. One was his failure to see or admit that the Newton of 1704 was not the Newton of 1687 or 1693. Berkeley's contention that no geometrical quantity can be exhausted by division is in consonance with the claim made by Zeno in his

Dichotomy or his *Achilles*.

In the *Analyst* Berkeley does not refer to Zeno, but according to Berkeley's argument, Achilles could not catch the tortoise. Nor can the modern reader agree with Berkeley in the claim that second or third fluxions are more mysterious than the first fluxion.

Berkeley experienced difficulty in conceiving fluxions as being proportional to the nascent increments or to the evanescent increments. Newton, himself, in his *Principia*, gave expression to the philosophical weakness of this explanation, for, strictly speaking, there is no first or prime ratio, nor is there a last or ultimate ratio. In his second reply to Berkeley Jurin defines a "nascent increment" in the Newtonian fashion as "less than any finite magnitude," also as "an increment just beginning to exist from nothing . . . but not yet arrived at any assignable magnitude how small so ever." Lagrange in a letter to Euler of November 24, 1759, said that he experienced trouble with Newton's exposition, since it considered the ratio of two quantities at the moment when they ceased to be quantities. Lagrange seems to have been convinced, says Jourdain, that the use of infinitesimals was rigorous and used both the infinitesimal method and the method of derived functions side by side, during his whole life. The question arises: did Berkeley believe that the calculus of fluxions was capable of rational exposition or not? Two noted mathematicians have vouchsafed opposite opinions on this point. Sir William Rowan Hamilton of quaternion fame says: "On the whole, I think that Berkeley persuaded himself that he was in earnest against fluxions, especially of order higher than the first, as well as against matter." To this De Morgan replied: "I have no doubt Berkeley knew that fluxions were sound enough." Berkeley himself said: "I have no controversy about your conclusions, but only about your logic and method." In view of the further fact that Berkeley in the *Analyst* advanced the theory of "Compensation of Errors" we incline to the opinion of De Morgan. The theory of "Compensation of Errors," we may add, was advanced independently by Lagrange and L. N. M. Carnot. According to Philip E. B. Jourdain this theory is found also in Maclaurin's *Fluxions*.

There are four other points in Berkeley's *Analyst* to which we desire to direct attention. First, his protestation against dropping quantities because they are comparatively very small. Jurin in his first reply argues in favor of the rejection of infinitesimals. In his second reply, after having received a castigation from Berkeley, Jurin says that this part of his ar-

gument was intended for popular consumption, for men such as one meets in London, who, when told that if Sir Isaac Newton were to measure the height of St. Paul's Church by fluxions he would be out not more than one tenth of a hair's breadth, and when further told that two books had been written in this controversy, would fly into a passion, would make reflections about "somebody's being overpaid," and would use expletives not fit for print.

Second, Berkeley's denial of the existence of infinitely small quantities is in conformity with the tenets of the recent school of Weierstrass and Georg Cantor.

Interesting is Berkeley's attack upon Newton's derivation of the moment or increment of a rectangle AB , as it is given in the *Principia*. Newton derives this moment by the difference

$$\left(A + \frac{1}{2}a\right) \left(B + \frac{1}{2}b\right) - \left(A - \frac{1}{2}a\right) \left(B - \frac{1}{2}b\right) \\ = Ab + Ba,$$

where a and b are assumed to be the increments of the sides. Berkeley argues with conviction that the increment of the rectangle AB is $bA + aB + ab$. Jurin takes the arithmetical mean of the increment $bA + aB + ab$ of the rectangle AB and of the decrement $bA + aB - ab$ of AB and obtains the desired true increment or "moment" as $aB + bA$. Sir William Rowan Hamilton sided with Berkeley against Newton on this point, but no eighteenth-century mathematician in England admitted the validity of Berkeley's criticism.

Last we come to the most fundamental of Berkeley's criticisms of Newton which centers upon what is called Berkeley's *lemma*: If in a demonstration an assumption is made, by virtue of which certain conclusions follow, and if afterward that assumption is destroyed or rejected, then all the conclusions that had been reached by the first assumption must also be destroyed or rejected. Berkeley applied this lemma to Newton's mode of deriving the fluxion of x^n as given in the *Quadrature of Curves* of 1704. Newton gives x a finite increment o , expands $(x + o)^n$ by the binomial formula, subtracts x^n and divides the remainder by o . He then lets o be zero and obtains the fluxion nx^{n-1} . Berkeley says that this reasoning is not fair or conclusive. "For when it is said, let the increment be nothing, the former supposition that the increment be something is destroyed and yet the expression got by that former supposition is retained." By Berkeley's lemma, this

is a false way of reasoning, “such as would not be allowed of in Divinity.”

It is interesting to observe that no British mathematician of the eighteenth century acknowledged the soundness of Berkeley’s *lemma* and its application.

Jurin, in his second reply to Berkeley, argues against the *lemma* thus: “You say that if one supposition be made, and be afterwards destroyed by a contrary supposition, then everything that followed from the first supposition is destroyed with it.” Not so, when the supposition and its contradiction are made at different times. “Let us imagine yourself and me to be debating this matter in an open field, . . . a sudden violent rain falls. . . we are all wet to the skin. . . it clears up. . . you endeavor to persuade me I am not wet. The shower, you say, is vanished and gone and consequently your wetness must have vanished with it.” The first recognition in England of the soundness of Berkeley’s lemma came in 1803 from Robert Woodhouse, who, in his *Principles of Analytical Calculation*, says that the methods of treating the calculus “all are equally liable to the objection of Berkeley, concerning the *fallacia suppositionis*, or the *shifting of the hypothesis*.” In finding the fluxion of x^n , the binomial expansion is effected “on the express supposition, that o is some quantity, if you take o equal to zero, the hypothesis is, as Berkeley says, shifted and there is a manifest sophism in the process.”

After Berkeley terminated his debate with the mathematicians, two mathematicians started a quarrel among themselves. Thus arose the second controversy on fluxions, which is comparatively little known.

Benjamin Robins, a self-educated mathematician, felt that Jurin had not entered a satisfactory defence of Newton, so Robins himself in 1735 published a tract entitled *A Discourse Concerning the Nature and Certainty of Sir Isaac Newton’s Method of Fluxions and of Prime and Ultimate Ratios*. Robins makes no reference to Berkeley or Jurin, or to their controversy. He lays the foundation of the calculus upon the concept of a limit. He speaks of a limit as a magnitude “to which a varying magnitude can approach within any degree of nearness whatever, though it can never be made absolutely equal to it.” Here for the first time is the stand taken openly, clearly, explicitly, that a variable can never reach its limit. From the standpoint of debating, this stand is a decided gain, but it is a gain made at the expense of generality. He descends to a very special type of variation which is not the variation encoun-

tered in ordinary mechanics; it is an artificial variation which does not permit Achilles to catch the tortoise. But this narrow concept of a limit nevertheless answers very well the needs of ordinary geometry. Robins’s tract is remarkable for clearness and soundness of exposition; it is a marked advance in that respect. The use of infinitely small quantities is rigidly excluded. The objections raised by Berkeley against *Fluxions* did not apply to Robins’s exposition. A long account of Robins’s *Discourse*, prepared by Robins himself, was published in a London monthly called *The Present State of the Republick of Letters*. In the next number of this monthly appeared an article by Jurin, under the pseudonym “Philalethes Cantabrigiensis,” in which he says that in his debate with Berkeley he adhered strictly to Newton’s language, but that some other defenders of Newton (meaning Robins) were guilty of departing from it. Jurin argues that the words *fiunt ultimo aequales* used by Newton in Lemma I of Book I in the *Principia*, mean that the quantities (the inscribed and circumscribed polygons) “at last become actually, perfectly, and absolutely equal”; in modern phraseology, the limit is reached. Several passages in the writings of Newton are examined and many illustrations are given. Robins prepared a reply to Jurin, and thus a controversy had gotten under way which threatened at one time to become endless. For two years there was a steady stream of articles in the *Republick of letters* and its successor *The Works of the Learned*. Pemberton entered the controversy during the second year on the side of Robins, but contributed nothing of value. About twenty articles were written, one of which filled one hundred and thirty-six pages. All articles taken together covered over seven hundred printed pages. They were attempts to ascertain what Newton’s ideas of fluxions and moments were, and whether Newton meant that a variable can reach its limit or cannot. And a good part of this material has escaped the attention of mathematical historians until now. The first few articles displayed care and ability, the later articles suffered in scientific value from the excessive heat of controversy. Jurin’s articles against Robins are superior to his articles against Berkeley. The debate is the most thorough discussion of the theory of limits carried on in England during the eighteenth century. It constitutes a refinement of previous conceptions. In my judgment Jurin’s interpretation of Newton was more nearly correct than that of Robins. The two disputants examined and reexamined every passage of Newton’s printed papers bearing on flux-

ions. Robins saw in Newton's condensed writings only variables which do not reach their limits; Jurin insisted that Newton permitted variables to reach their limits. Jurin admitted the calculus could be consistently founded upon Robins's idea of a limit, but he also insisted that Robins misrepresented Newton. Jurin's conceptions were quite broad for his time. He said: "Now whether a quantity or a ratio shall arrive at its limit or shall not arrive at it, depends entirely upon the supposition we make of the time during which the quantity or ratio is conceived constantly to tend or approach towards its limit." In other words, whether a variable reaches its limit or not is a matter of choice. We may impose conditions, so that the variable reaches its limit, or conditions under which it does not reach its limit. Thus Jurin was perhaps the first consciously to modify and generalize the limit concept. Modifications and generalizations of this have been going on ever since and are still in progress. A serious difficulty in permitting variables of the kind ordinarily arising in geometry to reach their limits lay in the fact that the imagination is not able to follow the variable through an infinity of steps that lead into the limit. The imagination exhausts itself in the effort. It is right here that Robins's variables which do not reach their limits had a great advantage. Jurin took great pains to devise illustrations of limit-reaching variables, intended to aid the imagination, though, as he admits, incapable of exhibiting the process "all the way." In one place Jurin says: "Since Mr. Robins is pleased to talk so much about straining our imagination, . . . let us see if we cannot find some plain and easy way of representing to the imagination that actual equality, at which the inscribed and circumscribed figures will arrive with each other, and with the curvilinear figure, at the expiration of the finite time." His procedure amounts to expressing the inscribed and circumscribed polygons as functions of the time, such that the limit is reached in a finite time.

It is interesting that toward the end of his long debate with Robins, Jurin begins to disavow infinitely small quantities. He brings out the difference between infinitesimals as variables, and infinitesimals as constants. He rejects all quantity "fixed, determinate, invariable, indivisible, less than any finite quantity whatsoever," but he usually admits somewhat hazily a quantity "variable, divisible, that, by a constant diminution, is conceived to become less than any finite quantity whatever, and at last to vanish into nothing."

Soon after the Berkeley onslaught, there appeared

nine British texts on fluxions, only one of which was of decidedly inferior type. None of these texts refer to the Jurin-Robins dispute. The latter was not widely noticed. The two thought-compelling publications that were widely read were Newton's *Quadrature of Curves* of 1704, and Berkeley's *Analyst*. The latter tract was always criticized by the mathematicians, yet always held in awe. These two tracts, together with Robins's *Discourse*, and Maclaurin's celebrated work on *Fluxions*, which appeared in 1742, mark the highest point of logical precision reached in England during the eighteenth century. All three of the great sections of the British Isles had contributed to this end: England through Newton and Robins, Ireland through Berkeley, and Scotland through Maclaurin. Maclaurin was familiar with the writings of the other three. He took the Greek demonstrative rigor as a model. In a biography of Maclaurin it is stated that several years before the publication of his fluxions, his demonstrations had been communicated to Berkeley and "Mr. Maclaurin had treated him with the greatest personal respect and civility; notwithstanding which, in his pamphlet on tar-water, he (Berkeley) renews the charge, as if nothing had been done" to remove the logical difficulties. Maclaurin avoided the use of infinitely small quantities, "an infinitely little magnitude being," as he expressed himself, "too bold a postulatium for such a science as geometry." He laid less stress upon the concept of a limit than did Robins and Jurin, and followed more closely the kinematical concepts of Newton. The term velocity had been the subject of dispute between Berkeley and Walton. Maclaurin perceived the difficulty of arguing that variable velocity is a physical fact. He defined the velocity of a variable motion as the space that *would be described* if the motion had continued uniform. He also quotes Barrow: Velocity is the "power by which a certain space may be described in a certain time" and then explains "power" by the consideration of "cause" and "effect" in a way that sounds odd in a work on fluxions. However, when we think of the Thomson-Dirichlet Principle we must acknowledge that the eighteenth century was not the only time when physical concepts were brought to the aid of mathematical theory. Apparently following Robins, Maclaurin's explanations imply that he does not encourage variables actually to reach their limits. Maclaurin secured his rigor of demonstration at a tremendous sacrifice. His work on *Fluxions* consists of seven hundred sixty-three pages; the first five hundred ninety pages do not contain the notation of

fluxions at all; the mode of exposition is rhetorical. This part deals with the derivation of the fluxions of different geometric figures, of logarithms, of trigonometric functions, also with the discussion of maxima and minima, asymptotes, curvature and mechanics, in a manner that the ancients might have adopted and with a verbosity of which the ancients are guilty. The consequence was that the work was not attractive reading. It was much praised and much neglected. Fifty-nine years elapsed before a second edition appeared. As we shall see, the book did not stop disputes on fluxions.

The middle and latter part of the eighteenth century were periods of mediocrity. There appeared a dozen books on fluxions, of which those of William Emerson and Thomas Simpson were the most noted. Both Emerson and Simpson were self-educated mathematicians, possessing the strength and the weakness usual with such preparation. Emerson returned to the use of infinitely small quantities, but a fluxion was defined as a velocity. This return to the use of infinitely small quantities is noticeable in several English texts of the second half of the century. An old lady once defended Calvinism by saying that if you took away her total depravity you took away her religion. There were mathematicians who believed that if you took away infinitely small quantities you took away all their mathematics. Simpson, in his text of 1750, which is a thorough revision of his text of 1737, avoids the use of infinitely small quantities. His definition of fluxion is as follows: "The magnitude by which any flowing quantity *would be* uniformly increased in a given time, with the generating celerity at any proposed position, or instant (was it from thence to continue invariable), is the fluxion of the said quantity at that position or instant." Substantially this definition of a fluxion was adopted later by Charles Hutton. Simpson dodges the word velocity, and remarks: "If motion in (or at) a point be so difficult to conceive that some have gone even so far as to dispute the very existence of motion, how much more perplexing must it be to form a conception, not only of the velocity of a motion, but also infinite changes and affections of it, in one and the same point, where all the orders of fluxions have to be considered." Simpson's definition and treatment of fluxions avoided the fictitious infinitesimals, as well as the perplexing term "velocity." Nevertheless, it did not enjoy security against attack, but was fiercely criticized in the *London Monthly Review*. The critic claimed that it is objectionable to define fluxion as the "magni-

tude by which any flowing quantity would be uniformly increased," for it was argued, that "in quantities *uniformly* generated, the *fluxion* must be the *fluent* itself, or else a part of it." It was claimed that Simpson's endeavor to exclude velocity "cannot be made intelligible without introducing velocity into it." "Again he mistakes the *effect* for the *cause*; for the thing generated must owe its existence to something, and this can only be the velocity of its motion, but it can never be the *cause itself*, as his definition would erroneously suggest." This obscure criticism of obscure points in Simpson's exposition initiated a third debate on fluxions which was carried on in the *Ladies Diary* and in ephemeral journals called the *Palladium*, the *Lady's Philosopher*, and the *Mathematical Exercises* (edited by John Turner). The debate was carried on between friends of Emerson on one side and friends of Simpson on the other. Emerson and Simpson do not themselves appear in the controversy. The friends of Emerson published in 1752 in London an anonymous pamphlet, entitled *Truth Triumphant or Fluxions for the Ladies, Showing the Cause to be Before the Effect, etc.*, which was criticized by the friends of Simpson as a "scurrilous pamphlet." It contains much that is foolish, a few passages eulogizing the works of Emerson, but also critical considerations which are of some interest and disclose the need of a more satisfactory arithmetical continuum. All in all this debate was carried on upon a much lower scientific plane than the former debates. The debaters represented the rank and file of mathematicians.

In the second half of the century several abortive attempts at arithmetization of the calculus were made. The most worthy of these attempts is due to John Landen, but his analysis is so complicated as to be prohibitive. Towards the latter part of the eighteenth century the efforts at rigorous exposition, which were so conspicuous in the years 1735–1742, slackened more and more. Colin Maclaurin was seldom read and John Robins was altogether forgotten. William Hales's discovery of Robins's *Discourse* in 1804 astonished him as would the discovery of a new work of Archimedes. The first edition of the *Encyclopedia Britannica*, 1771, permitted a "fluxion" to degenerate into an "increment" acquired in "less than any assigned time." The same article on fluxions appeared in the second edition (1779) and in the third edition (1797). In 1801 there was published in London Agnesi's *Analytical Institutions*, which many years earlier had been translated by John Colson from the Italian into English. How Colson's

conscience must have troubled him, when a fluxion stood out in his translation as something “infinitely small,” may be judged by the consideration that in 1736 he brought out an English translation of Newton’s *Method of Fluxions*. With Newton a fluxion always meant a finite velocity. We wonder what Robins and Maclaurin would have thought had they been alive in 1797 and 1801 and read these definitions. What horrible visions would these ghosts of departed quantities have brought to Bishop Berkeley had he been alive!

As we look back over the century we see that the eight years immediately following Berkeley’s *Analyst* were *eight great years*, during which Jurin, and especially Robins and Maclaurin made wonderful strides in the banishment of infinitely small quantities and the development of the concept of a limit. Both before and after that period of eight years, there existed in most writings of the eighteenth century in Great Britain, a mixture of Continental and British conceptions of the new calculus, a superposition of British symbols and phraseology upon the older Continental concepts. The result was a system, destitute of scientific interest. Newton’s notation was poor and Leibniz’s philosophy of the calculus was poor. That result represents the temporary survival of the least fit of both systems. The more recent international course of events has been in a diametrically opposite direction, namely, not to superpose Newtonian symbols and phraseology upon Leibnizian concepts, but, on the contrary, to superpose the Leibnizian notation and phraseology upon the limit-concept, as developed by Newton, Jurin, Robins, Maclaurin, d’Alembert and later writers.

About the opening of the nineteenth century more recent continental authors began to attract the attention of the English. Extensive accounts appeared in the *London Monthly Review* of Lagrange’s *Theory of Functions*, Lacroix’s *Differential Calculus*, Carnot’s *Reflexions on the Metaphysics of the Infinitesimal Calculus*. These texts were compared with English publications in a way not altogether favoring the English. Finally in 1805 Robert Woodhouse of Caius College, Cambridge, brought out his *Principles of Analytical Calculation* which contained many keen criticisms of both Continental and British mathematicians. Woodhouse is the first English mathematician who had a good word for Berkeley. He said: “I cannot quit this part of my subject without commenting on the *Analyst* and the subsequent pieces, as forming the most satisfactory controversial discussion in pure science that ever yet appeared:

into what perfection of perspicuity and logical precision the doctrine of fluxions may be advanced, is no subject of consideration; but view the doctrine as Berkeley found it, and its defects in metaphysics and logic are clearly made out. If for the purpose of habituating the mind to just reasoning . . . I were to recommend a book, it would be the *Analyst*.”

Woodhouse is the forerunner in Cambridge of Babbage, Peacock, and the younger Herschel, in the promotion of the principles of pure *D*-ism in opposition to the dot-age of the university.

As usually happens in reformations so here there was discarded and lost not only what was antiquated, but also what was meritorious. Robins’s *Discourse* of 1735, with its full and complete disavowal of infinitesimals and clear-cut, though narrow, conception of a limit was quite forgotten and d’Alembert’s definition was recommended and widely used in England. Now Robins and d’Alembert had the same conception of a limit. Both held the view that variables cannot reach their limits. However, there was one difference: Robins embodied this restriction in his definition of a limit; d’Alembert omitted it from his definition, but referred to it in his explanatory remarks.

Some of the eighteenth-century British conceptions possessed great merit. Perhaps no intuitional conceptions available in the study of the calculus are clearer and sharper than *motion* and *velocity*. These ideas offer even now great help in approaching the first study of the calculus. A second point of merit lay in the abandonment of the use of infinitely small quantities. Not all English authors of the eighteenth century broke away from infinitesimals, but those who did were among the leaders: Robins, Maclaurin, Simpson, Vince, and a few others. From the standpoint of rigor, the treatment of the calculus by these men was far in advance of the Continental. In Great Britain there was achieved in the eighteenth century in the geometrical treatment of fluxions that which was not achieved in the algebraical treatment until the nineteenth century. It was not until after the time of Weierstrass that infinitesimals were cast aside by mathematical writers on the Continent.

Judged by modern standards all eighteenth century expositions of the calculus, even the best British expositions, are defective. As pointed out by Landen and Woodhouse, there was an unnaturalness in founding the calculus upon *motion* and *velocity*. These notions apply in a real way only to dynamics. Moreover, not all continuous curves can be conceived as traceable by the motion of a point. The

notion of variable velocity is encumbered with difficulties. Then again, in all discussion of limits during the eighteenth century, the question of the *existence of a limit* of a given sequence was never raised. The word “quantity” was not defined; quantities were added, subtracted, multiplied and divided. Were these quantities numbers, or were they considered without reference to number? Both methods are possible. Which did British authors follow? No explicit answer to this was given. Our understanding of authors like Maclaurin, Rowe and others, is that in initial discussions such phrases as “fluxion of a curvilinear figure” are used in a non-arithmetical sense; the idea is purely geometrical. When later the finding of the fluxions of terms in the equations of curves is taken up, the arithmetical or algebraical conception is predominant. Rarely does a

writer speak of the difference between the two. Perhaps

His notions fitted things so well
That which was which he could not tell.

The theory of irrational number caused no great anxiety to eighteenth-century workers. Operations applicable to rational numbers were extended without scruple to a domain of numbers which embraced both rational and irrational. There was no careful exposition of the number system used. The modern theories of irrational number have brought about the last stages of what is called the *arithmetization* of mathematics. As now developed in books which aim at rigor the notion of a limit makes no reference to quantity and is a purely ordinal notion. Of this mode of treatment the eighteenth century never dreamed.

The Bernoullis and the Harmonic Series

WILLIAM DUNHAM

College Mathematics Journal 18 (1987), 18–23

Any introduction to the topic of infinite series soon must address that first great counterexample of a divergent series whose general term goes to zero — the harmonic series $\sum_{k=1}^{\infty} 1/k$. Modern texts employ a standard argument, traceable back to the great 14th-century Frenchman Nicole Oresme (see [3], p. 92), which establishes divergence by grouping the partial sums:

$$\begin{aligned} 1 + \frac{1}{2} &> \frac{1}{2} + \frac{1}{2} = \frac{2}{2}, \\ 1 + \frac{1}{2} + \left(\frac{1}{3} + \frac{1}{4}\right) &> \frac{2}{2} + \left(\frac{1}{4} + \frac{1}{4}\right) = \frac{3}{2}, \\ 1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \left(\frac{1}{5} + \frac{1}{6} + \frac{1}{7} + \frac{1}{8}\right) \\ &> \frac{3}{2} + \left(\frac{1}{8} + \frac{1}{8} + \frac{1}{8} + \frac{1}{8}\right) = \frac{4}{2}, \end{aligned}$$

and in general

$$1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{2^n} > \frac{n+1}{2},$$

from which it follows that the partial sums grow arbitrarily large as n goes to infinity.

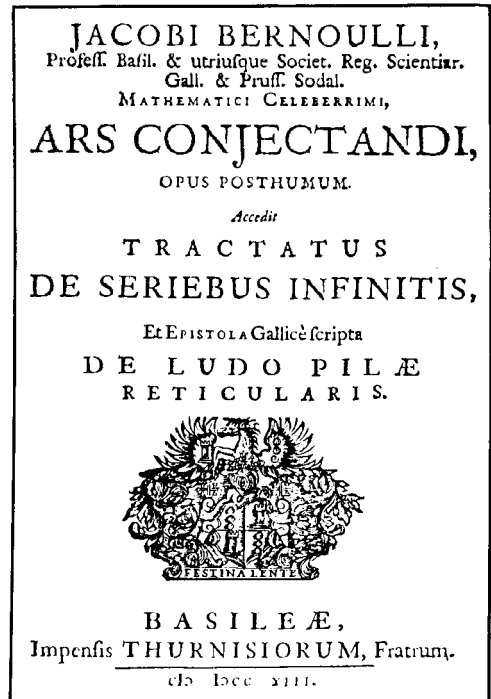
It is possible that seasoned mathematicians tend to forget how surprising this phenomenon appears to the uninitiated student—that, by adding ever more negligible terms, we nonetheless reach a sum greater than any preassigned quantity. Historian of mathematics Morris Kline ([5], p. 443) reminds us that this feature of the harmonic series seemed troubling, if not pathological, when first discovered.

So unusual a series could not help but attract the interest of the preeminent mathematical family of the 17th century, the Bernoullis. Indeed, in his 1689 treatise “Tractatus de Seriebus Infinitis,” Jakob Bernoulli provided an entirely different, yet equally ingenious proof of the divergence of the harmonic series. In “Tractatus,” which is now most

readily found as an appendix to his posthumous 1713 masterpiece *Ars Conjectandi*, Jakob generously attributed the proof to his brother (“Id primusprehendit Prater,”) the reference being to his full-time sibling and part-time rival Johann. While this “Bernoullian” argument is sketched in such mathematics history texts as Kline ([5], p. 444) and Struik ([6], p. 321), it is little enough known to warrant a quick reexamination.

The proof rested, quite unexpectedly, upon the *convergent* series

$$\frac{1}{2} + \frac{1}{6} + \frac{1}{12} + \frac{1}{20} + \cdots = \sum_{k=1}^{\infty} \frac{1}{k(k+1)}.$$



The modern reader can easily establish, via mathematical induction, that

$$\sum_{k=1}^n \frac{1}{k(k+1)} = \frac{n}{n+1},$$

and then let n go to infinity to conclude that

$$\sum_{k=1}^{\infty} \frac{1}{k(k+1)} = 1.$$

Jakob Bernoulli, however, approached the problem quite differently. In Section XV of *Tractatus*, he considered the infinite series

$$N = \frac{a}{c} + \frac{a}{2c} + \frac{a}{3c} + \frac{a}{4c} + \dots,$$

then introduced

$$P = N - \frac{a}{c} = \frac{a}{2c} + \frac{a}{3c} + \frac{a}{4c} + \frac{a}{5c} + \dots,$$

and subtracted termwise to get

$$\begin{aligned} \frac{a}{c} &= N - P \\ &= \left(\frac{a}{c} - \frac{a}{2c}\right) + \left(\frac{a}{2c} - \frac{a}{3c}\right) \\ &\quad + \left(\frac{a}{3c} - \frac{a}{4c}\right) + \dots \\ &= \frac{a}{2c} + \frac{a}{6c} + \frac{a}{12c} + \frac{a}{20c} + \dots \end{aligned} \tag{1}$$

Thus, for $a = c$, he concluded that

$$\frac{1}{2} + \frac{1}{6} + \frac{1}{12} + \frac{1}{20} + \dots = \frac{1}{1} = 1. \tag{2}$$

Unfortunately, Bernoulli’s “proof” required the subtraction of two divergent series, N and P . To his credit, Bernoulli recognized the inherent dangers in his argument, and he advised that this procedure must not be used without caution (“non sine cautela”). To illustrate his point, he applied the previous reasoning to the series

$$S = \frac{2a}{c} + \frac{3a}{2c} + \frac{4a}{3c} + \dots$$

and

$$T = S - \frac{2a}{c} = \frac{3a}{2c} + \frac{4a}{3c} + \frac{5a}{4c} + \dots$$

Upon subtracting termwise, he got

$$\frac{2a}{c} = S - T = \frac{a}{2c} + \frac{a}{6c} + \frac{a}{12c} + \frac{a}{20c} + \dots, \tag{3}$$

which provided a clear contradiction to (1).

Bernoulli analyzed and resolved this contradiction as follows: the derivation of (1) was valid since the “last” term of series N is zero (that is, $\lim_{k \rightarrow \infty} a/(kc) = 0$), whereas the parallel derivation of (3) was invalid since the “last” term of series S is non-zero (because $\lim_{k \rightarrow \infty} (k+1)a/(kc) = a/c \neq 0$). In modern terms, he had correctly recognized that, regardless of the convergence or divergence of the series $\sum_{k=1}^{\infty} x_k$, the new series $\sum_{k=1}^{\infty} (x_k - x_{k+1})$ converges to x_1 provided $\lim_{k \rightarrow \infty} x_k = 0$. Thus, he not only explained the need for “caution” in his earlier discussion but also exhibited a fairly penetrating insight, by the standards of his day, into the general convergence/divergence issue.

Having thus established (2) to his satisfaction, Jakob addressed the harmonic series itself. Using his brother’s analysis of the harmonic series, he claimed in Section XVI of *Tractatus*:

XVI. Summa seriei infinita harmonice progression- alium, $\frac{1}{1} + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \frac{1}{5}$ &c est infinita.

He began the argument that “the sum of the infinite harmonic series

$$\frac{1}{1} + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \frac{1}{5} \text{ etc.}$$

is infinite” by introducing

$$A = \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \frac{1}{5} + \frac{1}{6} + \frac{1}{7} + \dots,$$

which “transformed into fractions whose numerators are 1, 2, 3, 4 etc” becomes

$$\frac{1}{2} + \frac{2}{6} + \frac{3}{12} + \frac{4}{20} + \frac{5}{30} + \frac{6}{42} + \dots$$

Using (2), Jakob next evaluated:

$$\begin{aligned} C &= \frac{1}{2} + \frac{1}{6} + \frac{1}{12} + \frac{1}{20} + \dots = 1 \\ D &= \frac{1}{6} + \frac{1}{12} + \frac{1}{20} + \dots \\ &= C - \frac{1}{2} = 1 - \frac{1}{2} = \frac{1}{2} \\ E &= \frac{1}{12} + \frac{1}{20} + \dots \\ &= D - \frac{1}{6} = \frac{1}{2} - \frac{1}{6} = \frac{1}{3} \\ F &= \frac{1}{20} + \dots \\ &= E - \frac{1}{12} = \frac{1}{3} - \frac{1}{12} = \frac{1}{4} \\ &\vdots \\ &\vdots \end{aligned}$$

By adding this array columnwise, and again implicitly assuming that termwise addition of infinite series

References

1. Jakob Bernoulli, *Ars Conjectandi*, Basel, 1713.
2. Carl B. Boyer, *A History of Mathematics*, Princeton University Press, 1985.
3. C. H. Edwards, *The Historical Development of the Calculus*, Springer-Verlag, New York, 1979.
4. Leonhard Euler, *Opera Omnia* (1), Vol. 14 (C. Boehm and G. Faber, eds.), Leipzig, 1925.
5. Morris Kline, *Mathematical Thought from Ancient to Modern Times*, Oxford University Press, New York, 1972.
6. D. J. Struik (ed.), *A Source Book in Mathematics (1200–1800)*, Harvard University Press, 1969.
7. Translated from the Latin by Helen M. Walker, as noted in David E. Smith's *A Source Book in Mathematics*, Dover, New York, 1959, p. 271.

Leonhard Euler 1707–1783

J. J. BURCKHARDT

Mathematics Magazine 56 (1983), 262–273

Born in 1707, Leonhard Euler grew up in the town of Riehen, near Basel, Switzerland. Encouraged by his father, Paulus, a minister, young Leonhard received very early instruction from Johann I Bernoulli, who immediately recognized Euler's talents. Euler completed his work at the University of Basel at age 15, and at age 19 won a prize in the competition organized by the Academy of Sciences in Paris. His paper discussed the optimal arrangement of masts on sailing ships (*Meditationes super problemate nautico...*). In 1727 Euler attempted unsuccessfully to obtain a professorship of physics in Basel by submitting a dissertation on sound (*Dissertatio physica de sono*); however, this failure, in retrospect, was fortunate. Encouraged by Nicholas and Daniel, sons of his teacher Johann Bernoulli, he went to the St.

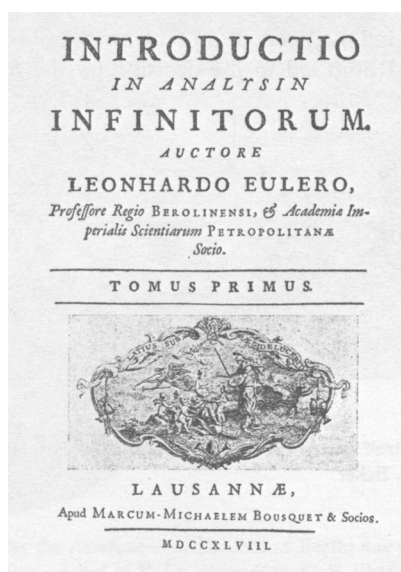
Petersburg Academy in Russia, a field of action that could accommodate his genius and energy.

In St. Petersburg Euler was met by compatriots Jacob Hermann and Daniel Bernoulli and so befriended the diplomat and amateur mathematician Christian Goldbach. During the years 1727–1741 spent there, Euler wrote over 100 scientific papers and his fundamental work on mechanics. In 1741, at the invitation of Frederick the Great, he went to the Akademie in Berlin. During his 25 years in Berlin, his incredible mathematical productivity continued. He created among other works, the calculus of variations, wrote the *Introductio in analysin infinitorum* (see Fig. 1), and translated and rewrote the treatise on artillery by Benjamin Robins.

Disputes with the Court led Euler in 1766 to ac-



Figure 1.



cept a very favorable invitation by Katherine II to return to St. Petersburg. There he was received in a princely manner, and he spent the rest of his life in St. Petersburg. Although totally blind, he wrote, with the help of his students, the famous *Algebra* and over 400 scientific papers; he left many unpublished manuscripts.

In recent decades, numerous important materials concerning Euler have been discovered in the archives in the Academy of Sciences of the USSR. It would seem that there is probably little chance of now discovering an unknown manuscript or something important about his life. Euler himself acknowledged the advantageous circumstances he found at the Academy. Judith Kh. Kopelevic notes, “Euler’s tombstone, erected by the Academy; his bust in the building of the Presidium of the Academy; the two-centuries-long efforts of the Academy to care for his enormous heritage and publish it—all these show clearly that Euler’s encounter with the Petersburg Academy of Sciences was a happy one for both sides.”

1 The legacy of Euler’s writings

Euler’s productivity is astonishing in its range of content and in the sheer volume of written pages. He wrote landmark books on the subjects of mathematical analysis, analytic and differential geometry, the calculus of variations, mechanics, and algebra. He published over 760 research papers, many of which won awards in competitions, and at his death left hundreds of unpublished works; even today there remain unpublished over 3,000 pages in notebooks. In view of this prodigious collection of written material, it is not surprising that soon after Euler’s death the task of surveying and publishing his works encountered extraordinary difficulty.

N. I. Fuss made efforts to publish more writings of the master, but only his son P.-H. Fuss succeeded (with the help of C. G. J. Jacobi) to generate interest among others, including Ostrogradskii. An enterprise in this direction was undertaken in Belgium (1838–1839), but failed after the publication of the fifth volume. In 1844, the Petersburg Academy decided on publication of the manuscripts, but this was not carried out. However, in 1849 the *Commentationes arithmeticae collectae*, edited by P.-H. and N. Fuss, were published; this contains, among others, the important manuscript *Tractatus de doctrina numerorum*.

The centennial of Euler’s death in 1883 rekindled interest in Euler’s works and in 1896 the most valuable preliminary to any complete publication appeared—the *Index operum Leonhardi Euleri* by J. G. Hagen. As the bicentennial of Euler’s birth neared, new life was infused into the project, which was thoroughly discussed by the academies of Petersburg and Berlin in 1903. Although the project was abandoned at this time, the celebrations of the bicentennial of Euler’s birth provided the needed impetus for the publication of the *Opera omnia*. The untiring efforts of Ferdinand Rudio led to the decision by the Schweizerische Naturforschende Gesellschaft [Swiss Academy of Sciences] in 1909 to undertake the publication, based on the list of Euler’s writings prepared by Gustaf Eneström (1910–1913). He lists 866 papers and books published by then. The financial side appeared assured through gifts and subscriptions. But the first World War led to unforeseen difficulties. We are indebted to Andreas Speiser for his efforts, which made it possible to continue the publication, and who overcame financial and publication difficulties so that at the start of World War II about one half of the project was completed. After the war, Speiser, succeeded by Walter Habicht, completed the series 1 (29 volumes), 2 (31 volumes) and 3 (12 volumes) of the *Opera omnia* except for a few volumes.

In 1947–1948 the manuscripts which had been loaned by the St. Petersburg Academy to the Swiss Academy of Sciences were returned to the archives of the Academy of Sciences of the USSR in Leningrad. Their systematic study was started under the supervision of the Academician V. I. Smirnov, with the goal of publishing a fourth series of the *Opera omnia*. As a first result, there appeared in 1965 a new edition of the correspondence between Euler and Goldbach, edited by A. P. Juskevich and E. Winter. In 1967, the Swiss Academy of Sciences and the Academia Nauk of the USSR formed an International Committee, to which was entrusted the publication of Euler’s correspondence in a series 4A, and a critical publication of the remaining manuscripts in a series 4B.

To mark the passage of 200 years since Euler’s death, a memorial volume has been produced by the Canton of Basel, *Leonhard Euler 1707–1783, Beiträge zu Leben und Werk*, edited by J. J. Burckhardt, E. A. Fellmann, and W. Habicht (Birkhäuser Verlag, Basel and Boston). From a contemporary point of view, this volume presents the insights of outstanding scientists on various aspects of Euler’s

achievements and their influence on later works. The complete list of essays and their authors appears at the end of this article. The memorial volume ends with a list, compiled by J. J. Burckhardt, of over 700 papers which are devoted to the work of Euler. It should be stressed that this is certainly an incomplete list, and it is hoped that it will lead to many additional listings which will then be published in an appropriate form. It is hoped that papers little known till now will receive the attention they deserve, and that this effort will lead to an improvement in the collaboration of scientists of all countries.

In the present article, we give a brief overview of the work of Euler. In order to include information from recently discovered work as well as the observations and insights of mode scholars, we draw freely from material found in the memorial volume.

2 Number theory

Euler had a passionate lifelong interest in the theory of numbers. Approximately one-sixth of his published work in pure mathematics is in this area; the same is true of the manuscripts left unpublished at his death. Although he had an active correspondence with Goldbach, he complained about the lack of response on the part of other contemporary mathematicians such as Huygens, Clairaut, and Daniel Bernoulli, who considered number theory investigations a waste of time, and were even unaware of Fermat's Theorem. (Forty years passed before Euler's investigations into Goldbach's problem were followed up by Lagrange.) André Weil has commented that if one were to distinguish between "theoretical" and "experimental" researchers, as is done for physicists, then Euler's constant preoccupation with number *theory* would place him among the former. But in view of his insistence on the "inductive" method of discovery of arithmetic truths, carrying out a wealth of numerical calculations for special cases before tackling the general question, one could equally well call him an "experimental" genius.

At the beginning of the eighteenth century—50 years after Fermat's death—the number-theoretical work of Fermat was practically forgotten. In a letter dated December 1, 1727, Christian Goldbach brought to Euler's attention Fermat's assertion that numbers of the form

$$2^{2^{p-1}} + 1, \quad p \text{ prime}$$

(i.e., 3, 5, 17, 257, ...) are also prime; this led Eu-

ler to a study of Fermat's works. His investigations included Fermat's Theorem and its generalizations, representations of numbers as sums of squares of polygonal numbers, and elementary quadratic forms.

In the decade between 1740 and 1750, Euler created the basis of a new theory which, until this day, has not essentially changed its character. The question which motivated this work was posed by Naudé on September 12, 1740, who asked Euler the number of ways in which a given integer can be represented as a sum of integers. For this problem, the "partitio numerorum," as well as for related problems, Euler found solutions by associating with a number-theoretic function its generating function, which can be investigated by analytical methods. Euler clearly understood the importance of his discovery. Although he had not found the proof of several central theorems of his theory, he incorporated the basic ideas and a few elementary but remarkable special results in his fundamental text in analysis, *Introductio in analysin infinitorum*. V. Scharlau comments, "Even today it is hard to imagine a more convincing and interesting introduction to this theory."

Euler used this theory in attempting to find a formula for prime numbers, where he considers the function $\sigma(n)$, the sum of all divisors of n . He obtained the formula

$$\sigma(p^k) = \frac{p^{k+1} - 1}{p - 1}, \quad \text{for } p \text{ prime}$$

from which the computation of $\sigma(n)$ follows. Euler also formulated the recursion rule for $\sigma(n)$,

$$\begin{aligned} \sigma(n) &= \sigma(n-1) + \sigma(n-2) \\ &\quad - \sigma(n-5) - \sigma(n-7) + \dots \end{aligned}$$

and observed its similarity to the one for $p(n)$, the number of partitions of n . In 1750, Euler brought these investigations to a conclusion by formulating the identity

$$\begin{aligned} &\prod_{i=1}^{\infty} (1 - x^i) \\ &= 1 + \sum_{m=1}^{\infty} (-1)^m (x^{\frac{1}{2}(3m^2-m)} + x^{\frac{1}{2}(3m^2+m)}) \end{aligned}$$

which is a cornerstone for all his related results.

Another interesting application of generating functions can be found in Euler's various investigations of "population dynamics," which probably originated in the years 1750–1755. Scharlau writes:

From today's point of view it is possibly not surprising that Euler found no additional results on generating functions; indeed it took many decades—almost a century—after the end of his activity before his achievements were substantially surpassed. It is remarkable how little attention was given to Euler's ideas by the mathematicians of the 18th and 19th centuries . . . There are very few mathematical theories whose character has changed so little since Euler's time as the theory of generating functions and the partitions of numbers.

Among the unpublished fragments of Euler's work (a total of about 3,000 pages, mainly bound in numbered notebooks) are over 1,000 pages which are devoted to number theory, mostly from the years 1736–1744 and 1767–1783. Euler's technique of investigation emerges clearly from these. After lengthy efforts which at times span many years, he reaches his results based on observations, tables, and empirically established facts.

G. P. Matvievskaia and E. P. Ozigova, who have perused these fragments, note that “the handwritten materials widen our views of Euler's activity in the field of number theory. The same holds for other directions of his research. The manuscripts enable us to recognize the sources of his mathematical discoveries.” A few examples serve to illustrate these points. On page 18 in notebook N131 is the problem of deciding whether a given integer is prime. The same notebook contains an entry about the origin of the zeta function, as well as the first mention of the theorem of four squares, to which Euler returns in notebook N132 (1740–1744). A particularly interesting entry in notebook N134 (1752–1755) contains Euler's formulation, a hundred years before Bertrand, of the “Bertrand postulate,” that there is at least one prime between any integer n and $2n$.

3 Analysis

Euler was occupied throughout his life with the concept of function; the treatises he produced in analysis were fundamental to the development of the modern foundations of analysis. As early as 1727 Euler had written a fifteen-page manuscript *Calculus differentialis*; it's interesting to compare this fledgling work with his later treatise *Institutiones calculi differentialis* (1755). Here Euler explains the calculus of finite differences of finite increments and considers calculations with infinitely small quantities. D.

Laugwitz, one of the contributors to the modern development of analysis through the adjoining of an infinity symbol Ω , remarks that anyone who reads this work, or Euler's *Introductio in analysin infinitorum* (1748), must be struck by the confidence with which Euler utilizes the calculus of both infinitely large and infinitely small magnitudes. Laugwitz indicates that it is possible to formulate Euler's ideas in the modern setting of nonstandard analysis, hence Euler receives a belated justification of his unorthodox techniques.

The richness and diversity of Euler's work in analysis can be seen by a brief summary of the book *Introductio in analysin infinitorum*. The first chapter discusses the definition of “function” which originated with Johann Bernoulli. In the second, Euler formulates the “fundamental theorem of algebra” and sketches a proof; he presents results on real and complex solutions of algebraic equations, a topic resumed in chapter 12 which deals with the decomposition of rational functions into partial fractions. The third chapter contains the so-called “Euler substitution” and the important replacement of a non-explicit functional dependence by a parametric representation. Particularly remarkable is Euler's strict theory of logarithms, and the consideration of the exponential function in chapter 6. Euler asserts that the logarithms of rationals are either rational or transcendental, a fact which was proved only two hundred years later. Weakly convergent series are considered in chapter 7, as well as the question of convergence of series and the relation between a function and its representation outside the circle of convergence. Subsequent chapters deal with transcendental functions and their representation as series or products. The starting point of Bernhard Riemann's investigation of the distribution of primes is in chapter 15, in the formula

$$\sum_n \frac{1}{n^x} = \prod_p \left(\frac{1}{1 - 1/p^x} \right)$$

which the summation extends over all positive integers and the product over all primes (see Fig. 2, left). In chapter 16 Euler turns to the new topic—rife with algebraic ideas—of *Partitione numerorum*, the additive decomposition of natural numbers (see Fig. 2, right). The developments of power series into infinite series found here were continued only by Ramanujan, Hardy and Littlewood. The expressions found here were later called theta functions, and used by Jacobi in the general theory of elliptic functions. The

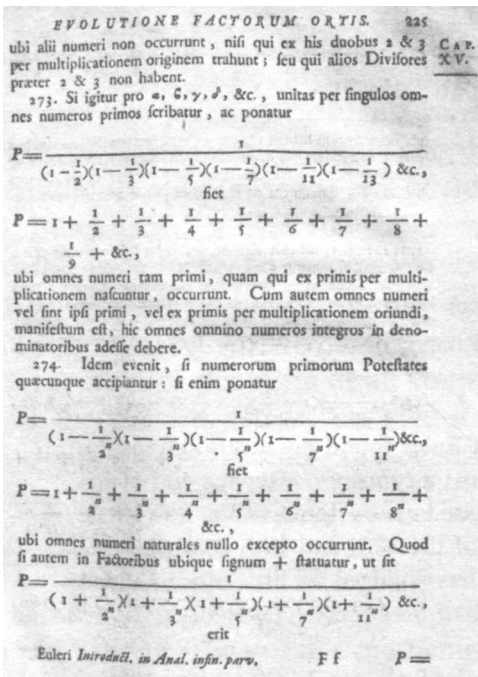
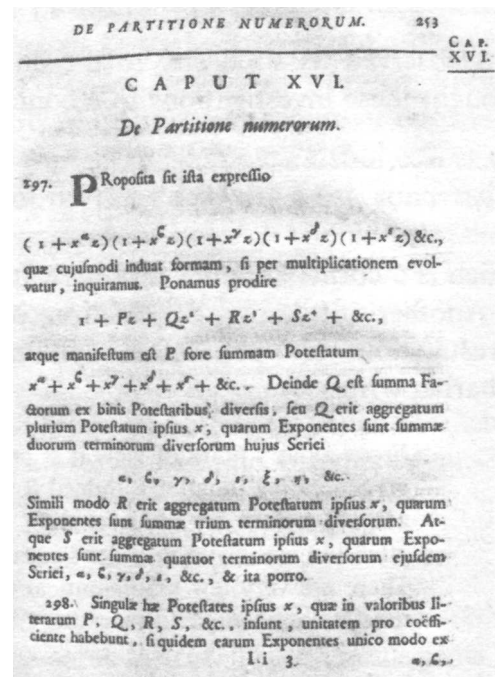


Figure 2.



last chapter, 17, deals with the numerical solution of algebraic equations, following Daniel Bernoulli.

A. O. Gelfond, whose essay in the memorial volume contains a deep analysis of the contents of *Introductio...*, interprets Euler’s ideas in modern terms and stresses the great relevance of this work, even to this day.

Euler’s interest in the theory of vibrating strings is legendary. In 1747 d’Alembert formulated the theory and the corresponding partial differential equation; this prompted Euler in 1750 to develop a solution, although restricted to the case in which the vibrations satisfy certain conditions. Euler’s friend Daniel Bernoulli contributed (about 1753) two remarkable articles, and presented the solution in the form of a trigonometric series. The problem is fittingly illuminated by Euler’s question “what is the law of the vibrating string if it starts with an arbitrary shape” and d’Alembert’s answer “in several cases it is not possible to solve the problem, which transcends the resources of the analysis available at this time.”

Euler has sometimes been criticized for seeming to ignore the concept of convergence in his freewheeling calculations. Yet in 1740, Euler gave an incomplete formulation of the criterion of convergence that later received Cauchy’s name. Euler’s last paper was completed in 1783, the year of his death; it contained

the germ of the concept of uniform convergence. His example was utilized by Abel in 1826.

After surveying the rich contributions to analysis made in Euler’s time, Pierre Dugac declares, “Euler and d’Alembert were the instigators of the most important work on the foundations of analysis in the nineteenth century.”

4 “Applied” mathematics (physics)

Euler’s investigations and formulations of basic theory in the areas of optics, electricity and magnetism, mechanics, hydrodynamics and hydraulics are among the most fundamental contributions to the development of physics as we know it today. Euler’s views on physics had an immediate influence on the study of physics in Russia; this grew out of his close relationship with the contemporary and most influential Russian scientist, M. V. Lomonosov, his several Russian students, and the publication of a translation (by S. J. Rumovskii) of his very popular “Letters to a German Princess” (see Fig. 3). The “Letters...” which had originated as lessons to the princess of Anhalt-Dessau, niece of the King of Prussia, during Euler’s years in Berlin, served as the first encyclopedia of physics in Russia. A. T. Grigorjan and V. S.

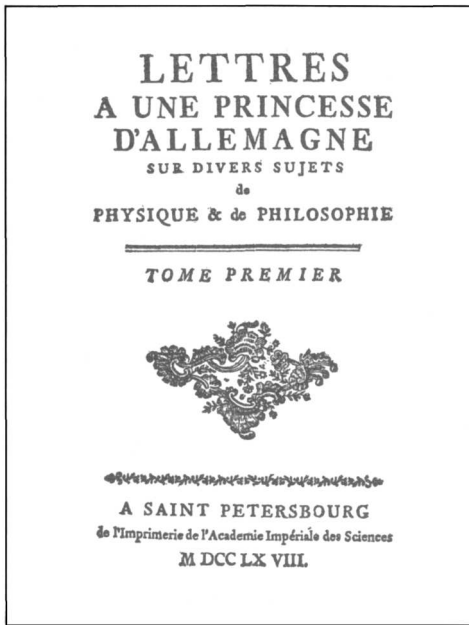


Figure 3.

Kirsanov have noted that the physicist N. M. Speranskii, a noted statesman and author of a physics book (1797), used to read to his students sections from Euler's "Letters. . ."

B. L. van der Waerden, in discussing Euler's justification of the principles of mechanics, has asked, "What did Euler mean by saying that in the computation of the total moment of all forces, the inner forces can be neglected because 'les forces internes se détruisent mutuellement'?" He points out that in order to answer that question it is important to know Euler's concept of solids, fluids, and gases. Are they true continua, or aggregates of small particles? The answer can be found in Euler's letters #69 and #70 to a German princess. He does not consider water, wool and air as true continua, but assumes that they consist of separate particles. However, in hydrodynamics, Euler treats liquids and gases as if they were continua. Euler is well aware that this is only an approximation.

A study of the published works of Daniel and Johann Bernoulli, as well as Euler's unpublished works (in particular, Euler's thick notebook from 1725–1727), by G. K. Mikhailov, gives some new and surprising insights into Euler's contributions to the development of theoretical hydraulics. Mikhailov states:

It is generally known that the creation of the foundations of modern hydrodynamics of ideal

fluids is one of the fruits of Euler's scientific activity. Less well known is his role in the development of theoretical hydraulics, that is, as usually understood, the hydrodynamic theory of fluid motion under a one-dimensional flow model. Traditionally—and with good reason—it is assumed that the foundations of hydraulics were developed by Daniel and Johann Bernoulli in their works published between 1729 and 1743. In fact, during the second quarter of the eighteenth century Euler did not publish even a single paper on the elements of hydraulics. The central theme of most of the recent historical-critical studies on the state of hydraulics in that period is the determination of the respective contributions of Daniel and of Johann Bernoulli. But Euler understood, all this time, just beyond the curtain of the stage on which the action was taking place, although almost no contemporary was aware of that.

Euler's work on the theory of ships culminated in the publication of *Scientia navalis seu tractatus de construendis ac dirigendis navibus*, published in 1749. Walter Habicht notes the fundamental importance of this treatise:

Following the *Mechanica sive motus scientia analytice exposita* which appeared in 1736, it [the *Scientia navalis*. . .] is the second milestone in the development of rational mechanics, and to this day has lost none of its importance. The principles of hydrostatics are presented here, for the first time, in complete clarity; based on them is a scientific foundation of the theory of shipbuilding. In fact, the topics treated here permit insights into all the related developments in mechanics during the eighteenth century.

Although Euler's intense interest in the science of optics appeared before he was 30 and remained with him almost to his death, there is still no monographic evaluation of his contributions to the wide field of physical and geometrical optics. Part of Euler's work is best described by Habicht:

In the second half of his life, from 1750 on and throughout the sixties, Leonhard Euler worked intensively on problems in geometric optics. His goal was to improve in several ways optical instruments, in particular, telescopes and microscopes. Besides the determination of the enlargement, the light intensity and the field of view, he was primarily interested in the devi-

ations from the point-by-point imaging of objects (caused by the diffraction of light passing through a system of lenses), and also in the even less tractable deviations which arise from the spherical shape of the lenses. To these problems Euler devoted a long series of papers, mainly published by the Berlin academy. He admitted that the computational solution of these problems is very hard. As was his custom, he collected his results in a grandly conceived textbook, the *Dioptrica* (1769–1771) (see Fig. 4). This book deals with the determination of the path of a ray of light through a system of diffracting spherical surfaces, all of which have their centers on a line, the optical axis of the system. In a first approximation, Euler obtains the familiar formulae of elementary optics. In a second approximation he takes into account the spherical and chromatic aberrations. After passing through a diffracting surface, a pencil of rays issuing from a point on the optical axis is spread out in an interval on the optical axis; this is the so-called “longitudinal aberration.” Euler uses the expression “espace de diffusion.” If the light passes through several diffracting surfaces, the “espace de diffusion” is determined using a principle of superposition.

Euler had great expectations for his theory, and believed that using his recipes, the optical instruments could be brought to “the highest degree of perfection.” Unfortunately, the practical realization of his systems of lenses did not yield the hoped-for success. He searched for the causes of failure in the poor quality of the lenses on the one hand, and also in basic errors in the laws of diffraction which were determined experimentally in a manner completely unsatisfactory from a theoretical point of view. Because of the failure of his predictions, Euler’s *Dioptrica* is often underrated.

Habicht notes that Euler’s theory can be modified to obtain the general imaging theories developed in the nineteenth century. The crucial gap in Euler’s treatment consists in neglecting those aberrations which are caused by the distance of the object and its images from the optical axis; with modification it is possible to determine the spherical aberration errors of the third order directly from Euler’s formulas.

A responsible evaluation of Euler’s contributions to optics will be possible only after Euler’s unpublished letters and manuscripts are edited and made

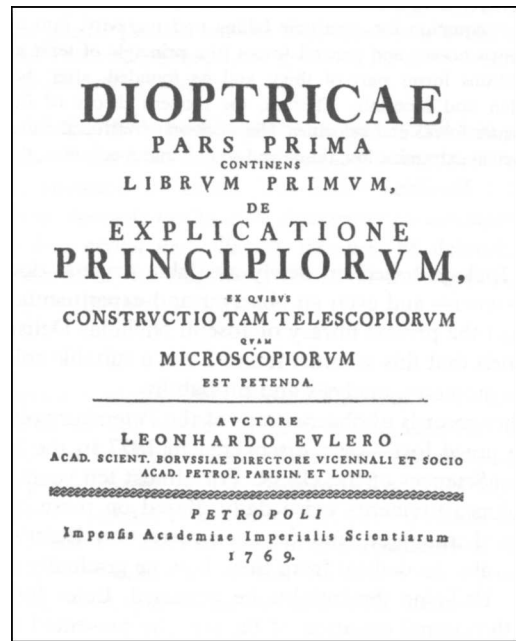


Figure 4.

generally accessible. E. A. Fellmann provides an example of Euler’s method which helps to place Euler’s contribution in a historic context. The problem of diffraction in the atmosphere is one which was first seriously considered by Euler:

He began by deriving a very general differential equation; naturally, it turned out not to be integrable—it would have been a miracle had that not happened. Then he searched for conditions which make a solution possible, and finally he solved the problem in several cases under practically plausible assumptions.

Euler frequently expressed the opinion that the phenomena in optics, electricity and magnetism are closely related (as states of the ether), and that therefore they should receive simultaneous and equal treatment. This prophetic dream of Euler concerning the unity of physics could only be realized after the construction of bridges (experimental as well as theoretical) which were missing in Euler’s time. These were later built by Faraday, W. Weber and Maxwell.

Euler was deeply influenced by the work of scientists who preceded him as well as by the work of his contemporaries. This is perhaps best illustrated by his role in the development of potential theory. He acknowledges the influence of the work of Leibniz, the Bernoullis, and Jacob Hermann, whose

work he had studied in his days in Basel to 1727. In the decade 1730–1740, the contemporaries Euler, Clairaut and Fontaine all were active in developing the main ideas that would lead to potential theory: the geometry of curves, the calculus of variations, and the study of mechanics. By 1752 Euler's work on fluid mechanics *Principia motus fluidorum* was complete. A summary of his contributions to potential theory is given by Jim Cross:

He helped, with Fontaine and Clairaut, to develop a logical, well-founded calculus of several variables in a clear notation; he transformed, with Daniel Bernoulli and Clairaut, the Galileo-Leibniz energy equation for a particle falling under gravity, into a general principle applicable to continuous bodies and general forces (the principle of least action with Daniel Bernoulli and Maupertuis forms part of this); and he founded, after the attempts of the Bernoullis, d'Alembert, and especially Clairaut, the modern theory of fluid mechanics on complete differentials for forces and velocities. His work was fruitful: the theories of Lagrange grew from his writings on extremization, fluids and sound, and mechanics; the work of Laplace followed.

5 Astronomy

Research by Nina I. Nevskaja based on newly available original documents justifies calling Euler a professional astronomer—and even an observer and experimental scientist. Five hundred books and manuscripts from the private library of Joseph Nicholas Delisle have recently come to light and from these one finds that this scientist found Euler a suitable collaborator and valued his knowledge in spherical trigonometry, analysis and probability.

It was a surprise when the records of observations of the Petersburg observatory during its first 21 years—which were presumed lost—were discovered in 1977 in the Leningrad branch of the archives of the Academy of Sciences of the USSR. For almost ten years, Euler was among those who were regularly taking measurements twice daily. Based on these observations, Delisle and Euler computed the instant of true noon, and the noon correction. Euler's entries were so detailed and numerous that it is possible to deduce from them how he gradually mastered the methods of astronomical observations. Utilizing the insights he obtained, Euler found a simple method

of computing tables for the meridional equation of the sun; he presented it in the paper *Methodus computandi aequationem meridiei* (1735).

Euler was fascinated by sunspots; his notes from this period contain enthusiastic comments on his observations. The computation of the trajectories of the sunspots by Delisle's method can be considered the beginning of celestial mechanics. The archives also disclose that Euler helped Delisle by working out analytical methods for the determination of the paths of comets.

A little-noted field of Euler's activities, the theory of motion of celestial bodies, is documented by Otto Volk. Euler's first paper, based on generally formulated differential equations of mechanics, is entitled *Recherches sur le mouvement des corps célestes en général* (1747). Using the tables of planets computed by Thomas Street from the pure Keplerian motion of planets around the sun, Euler discusses in Sections 1 to 17 the observed irregularities. In Section 18 he formulates the differential equations of mechanics, and obtains the solution

$$r = a(1 + e \cos v) = \frac{a(1 - e^2)}{1 - e \cos \phi}$$

in which r is the radius, v is the eccentric anomaly and ϕ is the true anomaly, while e and a are constants. This is a regularization of the so-called inverse problem of Newton. Later, Euler obtains a trigonometric series for ϕ ; such Fourier series are the basis of his computation of perturbations. This is the topic treated in detail in the prize proposal to the Paris Academy, *Recherches sur la question des inégalités du mouvement de Saturne et de Jupiter, sujet proposé pour le prix de l'année 1748*. In it Euler uses, for the first time, Newton's laws of gravitation to compute the mutual perturbations of planets.

In his paper *Considerationes de motu corporum coelestium* (1764), Euler is the first to begin considering the three-body problem, under certain restrictions. Euler notes the intractability of the problem:

There is no doubt that Kepler discovered the laws according to which celestial bodies move in their paths, and that Newton proved them—to the greatest advantage of astronomy. But this does not mean that the astronomical theory is at the highest level of perfection. We are able to deal completely with Newton's inverse-square law for two bodies. But if a third body is involved, so that each attracts both other bodies, all the arts of analysis are insufficient . . . Since the solution of the general problem of

three bodies appears to be beyond the human powers of the author, he tried to solve the restricted problem in which the mass of the third body is negligible compared to the other two. Possibly, starting from special cases, the road to the solution of the general problem may be found. But even in the case of the restricted problem the solution encounters difficulties so great that the author has to admit to have spent much effort in vain attempts at solution.

Euler's investigation of the three-body problem was noted only at a later date; the linear solutions to the equation of the fifth degree were (and sometimes still are) called "Lagrange's solutions," without any mention of Euler. But Euler achieved fame through his theory of perturbations, presented in *Nouvelle méthode de déterminer les dérangemens dans le mouvement des corps célestes, causé par leur action mutuelle*. By iteration he determined, for the first time, the perturbations of the elements of the elliptical paths, and then applied this method to determine the motion of three mutually attracting bodies.

6 Correspondence

The circle of contemporary scholars who were influenced by and in turn, influenced, Euler's investigations was as wide as one could imagine in the eighteenth century. His voluminous correspondence testifies to the fruitful interaction between scientists through queries, conjectures, critical comments, and praise. Some of the correspondence has been published previously in collected works; a standard reference is the collection *Correspondance Mathématique et Physique*, edited by N. Fuss and published in 1843 by the Imperial Academy of Science, St. Petersburg. New discoveries and more complete information have produced recently published collections. The publication in 1965 of the correspondence between Euler and Christian Goldbach has been mentioned earlier.

It is significant that the first volume, A1, published in the fourth series of Euler's *Opera omnia*, contains a complete list of all existing letters to and from Euler (about 3,000), together with a summary of their contents. Volume A5 of this series (1980), edited by A. P. Juskevic and R. Taton, contains Euler's correspondence with A. C. Clairaut, J. d'Alembert, and J. L. Lagrange.

The correspondence between Euler and Lagrange from 1754 to 1775 gives valuable testimony to the

development of personal relations between two of the most important scientists of that time. The letter exchange begins with a letter from the 18-year-old Lagrange, who lived in Turin, containing a query in which he mentions the analogy in the development of the binomial $(a+b)^m$ and the differential $d^m(xy)$. Mathematically isolated, Lagrange expresses his admiration for Euler's work, particularly in mechanics. Especially significant is the second letter to Euler (1755). In it Lagrange announces, without details, his new methods in the calculus of variations; Euler at once notes the advantage of these methods over the ones in his *Methodus inveniendi lines curvas maximi minimive proprietate gaudentes* (1744), and heartily congratulates Lagrange. In 1756 Lagrange develops the differential calculus for several variables and investigates, for the first time, minimal surfaces. After an interruption of three years, Lagrange continues the correspondence by sending his work *La nature et la propagation du son*, and we find interesting discussions on the problem of vibrating strings, which had been carried on since 1749 between d'Alembert, Euler and Daniel Bernoulli.

After a lengthy pause, Euler resumes the correspondence. The first letter (1765) concerns the discussion with d'Alembert on vibrating strings, and the librations of the moon. In a second, Euler tells Lagrange that he has been granted permission by Friedrich II to return to Petersburg, and is attempting to have Lagrange come there. In later correspondence, the emphasis is on questions in the theory of numbers and in algebra. Pell's equation $x^2 - ay^2 = b$; and in particular $p^2 - 13q^2 = 101$, are discussed. Other topics deal with arithmetic, questions concerning developable surfaces, and the motion of the moon.

In 1770 Lagrange writes of his plan to publish Euler's *Algebra* in French, and to add to it an appendix; the published book is mailed on July 13, 1773. The last of Euler's letters, dated March 23, 1775, is remarkable by the exceptionally warm congratulations for Lagrange's work, especially about elliptic integrals. It may be conjectured that this was not the end of the correspondence, but unfortunately no additional letters have survived.

7 Postscript

This overview of Euler's life and work touches only a small part of the wealth of material to be found in the scholarly essays in the Basel memorial volume.

In addition to careful and detailed analysis of many of Euler's scientific and mathematical achievements, these chapters contain new information on all aspects of Euler's private and academic life, his family, his philosophical and religious views, and the fabric of his life and work at the St. Petersburg Academy. In view of the overwhelming volume and diversity of Euler's work, it may never be possible to produce a comprehensive scientific biography of his genius. It is to be hoped that these newest contributions to the study of his life and work will provide impetus for further study and publication of many of the yet unpublished papers which are the unknown legacy of this mathematical giant.

Reference

Leonhard Euler 1707–1783, Beiträge zu Leben und Werk, Gedenkband des Kantons Basel-Stadt, edited by J. J. Burckhardt, E. A. Fellmann, and W. Habicht, Birkhäuser Verlag, Basel, 1983.

Table of Contents

Emil A. Fellmann (Basel, CH), Leonhard Euler — Ein Essay über Leben und Werk
 Aleksander O. Gelfond (1906–1968) (Moskau, UdSSR), Über einige charakteristische Züge in den Ideen L. Eulers auf dem Gebiet der mathematischen Analysis und in seiner 'Einführung in die Analysis des Unendlichen'
 André Weil (Princeton, USA), L'oeuvre arithmétique d' Euler
 Winfried Scharlau (Munster, BRD), Eulers Beiträge zur *partitio numerorum* und zur Theorie der erzeugenden Funktionen
 Galina P. Matvievskaia/Helena P. Ozigova (Taskent-Leningrad UdSSR), Eulers Manuskripte zur Zahlentheorie
 Adolf P. Juskevich (Moskau, UdSSR), L. Euler's unpublished manuscript *Calculus Differentialis*
 Pierre Dugac (Paris, F) Euler, d'Alembert et les fondements de l'analyse
 Detlef Laugwitz (Darmstadt, BRD), Die Nichtstandard-Analysis: Eine Wiederaufnahme der Ideen und Methoden von Leibniz und Euler

Isaac J. Schoenberg (Madison, USA), Euler's contribution to cardinal spline interpolation: The exponential Euler splines
 David Speiser (Louvain, Belgien), Eulers Schriften zur Optik, zur Elektrizität und zum Magnetismus
 Gleb K. Mikhailov (Moskau, UdSSR), Leonhard Euler und die Entwicklung der theoretischen Hydraulik im zweiten Viertel des 18. Jahrhunderts
 Walter Habicht (Basel, CH), Einige grundlegende Themen in Leonhard Eulers Schiffstheorie
 Bartel L. van der Waerden (Zürich, CH), Eulers Herleitung des Drehimpulssatzes
 Walter Habicht (Basel, CH), Betrachtungen zu Eulers Dioptrik
 Emil A. Fellmann (Basel, CH), Leonhard Eulers Stellung in der Geschichte der Optik
 Jim Cross (Melbourne, Australia), Euler's contributions to Potential Theory 1730–1755
 Otto Volk (Würzburg, BRD), Eulers Beiträge zur Theorie der Bewegungen der Himmelskörper
 Nina I. Nevskaja (Leningrad, UdSSR), Leonhard Euler und die Astronomie
 Judith Kh. Kopelevic (Leningrad, UdSSR), Leonhard Euler und die Petersburger Akademie
 Asot T. Grigor'jan and V. S. Kirsanov (Moskau, UdSSR), Euler's Physics in Russia
 Ivor Grattan-Guinness (Barnet, GB), Euler's Mathematics in French Science, 1795–1815
 René Taton (Paris, F), Les relations d'Euler et de Lagrange
 Pierre Speziali (Geneve, CH), Leonard Euler et Gabriel Cramer
 Roger Jaquel (Mulhouse, F), Leonard Euler, son fils Jean-Albrecht et leur ami Jean III Bernoulli
 Wolfgang Breidert (Karlsruhe, BRD), Leonhard Euler und die Philosophie
 Michael Raith (Riehen bei Basel, CH), Der Vater Paulus Euler. Beiträge zum Verständnis der geistigen Herkunft Leonhard Eulers
 René Bernoulli (Basel, CH), Leonhard Eulers Augenkrankheiten
 Kurt-Reinhard Biemann (Berlin, DDR), Aus der Vorgeschichte der Euler-Werkausgabe
 Johann Jakob Burckhardt (Zürich, CH), Die Eulerkommission der Schweizerischen Naturforschenden Gesellschaft. — Ein Beitrag zur Editions-geschichte
 Johann Jakob Burckhardt (Zürich, CH), Euleriana-Verzeichnis des Schrifttums über Leonhard Euler

The Number e

J. L. COOLIDGE

American Mathematical Monthly 57 (1950), 591–602

1 The Greek beginning

The distinguished American mathematician, Benjamin Peirce, was wont to find all of analysis in the equation

$$i^{-i} = \sqrt{e^\pi}.$$

In fact, he had his picture taken in front of a blackboard on which this mystic formula, in somewhat different shape, was inscribed. He would say to his hearers, “Gentlemen, we have not the slightest idea of what this equation means, but we may be sure that it means something very important.”

With regard to the symbols which appear in this charm, there is a vast literature connected with π ; and i , when written $\sqrt{-1}$, leads into the broad field of analysis in the complex domain; but it seems surprisingly difficult to find a connected account of e .

I think we may make a fair beginning with the twelfth proposition of the Second Book of Apollonius’ *Conics*, which tells us that if from a point on a hyperbola lines be drawn in given directions to meet the asymptotes, the product of the two distances is independent of the position of the point chosen on the curve. This theorem is more general than we shall need to arrive at the number e and it is not original with Apollonius. Let us confine ourselves to the very special case where the hyperbola is rectangular, and we draw to each asymptote a line parallel to the other. When x and y are distances, we may write

$$xy = 1. \quad (1)$$

It is intriguing to inquire who first discovered the theorem which leads to this equation. In the commentary of Eutocius on the *Sphere and Cylinder of Archimedes* [1], we come to a discussion of the classical problem of inserting two mean proportionals between two given lengths. In one solution, which

he labels “ut Menaechmus,” we have what amounts to the equations

$$\begin{aligned} a/x &= x/y = y/b; \\ y^2 &= bx; \quad xy = ab. \end{aligned} \quad (2)$$

He goes on to seek the intersection of a parabola and a hyperbola.

Eutocius’ statement would place the theorem very early in the history of the conics, for Menaechmus is usually regarded as the discoverer or inventor of these curves, although this ascription is by no means certain. Allman writes [2], “It is much to be regretted that the two solutions of Menaechmus have not been transmitted to us in their original form. That they have been altered either by Eutocius or by some author whom he followed appears not only in the employment in these solutions of the terms parabola and hyperbola, as has frequently been pointed out, but more from the fact that the language used in them is, in character, altogether that of Apollonius.” A similar doubt is shown in Loria [3]. On the other hand, Heath is perfectly definite on this point; he states, “This property in the particular case of the rectangular hyperbola was known to Menaechmus” [4].

But there is another reason for doubting the ascription to Menaechmus, aside from the linguistic objection. The classical Greek discussion of the conics always corresponds to our analysis when the axes are a tangent and the diameter through the point of contact, and with these data proofs are not simple. Heath, following Zeuthen, shows the fact that the hyperbola can be written immediately in the form (1) if we start with a technique like ours, that is, when the axes are a pair of conjugate diameters [5]. That is perfectly true, but the Greeks made surprisingly little study of the conics when expressed in

this form more familiar to us; Apollonius comes to it quite late. It seems to me altogether doubtful that the first discoverer of the curves should have been able to make the transition.

2 Grégoire de St. Vincent

If we grant that the Greek mathematicians, perhaps Menaechmus, were familiar with the fundamental property of the rectangular hyperbola expressed in (1), what has this to do with e ? We must look ahead some two thousand years to that original writer whose name appears at the head of this paragraph. In 1647, he published his fundamental *Prologomena a Santo Vincento, Opus geometricum quadraturae circuit et sectionum conii*. This I have not seen in its original form, but the content is given at great length by Bopp in [6]. Here is the general scheme. We take the hyperbola

$$xy = 1. \tag{3}$$

On the x -axis we take n equivalent rectangles whose bases are

$$P_0P_1, P_1P_2, \dots, P_{n-1}P_n,$$

while each has an upper vertex on the curve Q_i . Then,

$$P_0P_1 \cdot P_0Q_0 = P_1P_2 \cdot P_1Q_1 = P_2P_3 \cdot P_2Q_2 = \dots,$$

and

$$\frac{P_0Q_0}{P_1Q_1} = \frac{P_1P_2}{P_0P_1}; \quad \frac{P_1Q_1}{P_2Q_2} = \frac{P_2P_3}{P_1P_2}, \tag{4}$$

but

$$OP_0 \cdot P_0Q_0 = OP_1 \cdot P_1Q_1 = OP_2 \cdot P_2Q_2 = \dots,$$

so that

$$\frac{OP_0}{OP_1} = \frac{P_1Q_1}{P_0Q_0} = \frac{P_0P_1}{P_1P_2} = \frac{OP_1}{OP_2},$$

by composition. If

$$OP_1 = \rho OP_0, \text{ then } OP_j = \rho^j OP_0. \tag{5}$$

St. Vincent even treats the case where OP_0 and OP_j are incommensurable, but we need not follow him here.

The importance of this equation was early recognized, because of its connection with logarithms which were based on the relation of arithmetical and geometrical series. There is a good deal to be said in

favor of the thesis that the credit for relating the rectangular hyperbola with logarithms is due to Sarasa. I have not seen his work, but like Cantor, I rely on Kästner. In 1649, Sarasa published *Solutio problematis a R. P. Marino Mersenne propositi*. This was concerned with the problem: Given three positive quantities and the logarithms of two of them, find the logarithm of the third. Kästner writes [7], "Zu ihrer Beantwortung brang Sarasa drey Saetze aus des Gregorius Buche von der Hyperbel bey, die betreffen Flaechen der Hyperbel an der Aysmptoten, Sarasa erinnert wie das mit Logarithmen zusammenhangt." Cantor's view is similar [8]; he states, "Mit andern Worten, Gregorius hatte das Auftreten von Logarithmen bei der erhahnten Flächenraumen erkannt, wen auch nicht mit Namen genannt. Letzteres that Sarasa, und darin liegt das wirkliche Verdienst seiner Stratschrift."

A contrary view is expressed by Charles Hutton [9] in the words, "As to the first remarks on the analogy between logarithms and hyperbolic spaces, it having been shown by Gregory St. Vincent . . . that if an asymptote be divided into parts in geometrical progression, and from the points of division ordinates be drawn parallel to the other asymptote, they will divide the space between the asymptote and the curve into equal portions, from hence it was shown by Mersenne, that by taking continual sums of these parts there would be obtained areas in arithmetical progression which therefore were analogous to a system of logarithms."

This may be true, but I must point out that whereas St. Vincent published the work referred to above in 1647, Mersenne died in the middle of 1648, and the dates of all of his mathematical writings which I have seen were much earlier. However, St. Vincent's work was certainly well observed. We find Wallis writing in 1658 to Lord Brouncker [10], "Sumptis (in Asymptoto) rectis NH, NI, NK, NQ, NL, NM geometricre proportionalibus, in punctis H, I, K, Q, L, M , ducantur rectae parallelae alteri Asymptoto, spatium Hyperbolicum $ABHM$ in quinque partes dividi ostendit Gregor de Sancto Vincento (si memini) decimo."

3 The introduction of logarithms

The actual word logarithm occurs again in an account of Gregory's *Vera circuli et hyperbolae quadratura*, which was published in Padua in 1667

and laid before the Royal Society [11]. Here we read, “And lastly by the same method he calculates both the logarithm of any natural number, and, vice versa, the natural number of any given logarithm.” Perhaps the wisest word on the subject has been pronounced by the kindly old writer Montucla [12], “Au reste la découverte de cette propriété est revendiquée par divers autres géomètres.” Among these I surely must mention Christian Huygens, who acknowledges the work of St. Vincent, even though he does not claim for himself the discovery of the relation between the hyperbola and logarithms. This is admirably set forth in [13], first in a French account, then Huygens’ own Latin. He finds the areas bounded by the x axis, which is an asymptote, the curve and ordinates. Two such areas terminating by the same ordinate of 1 are

$$\frac{\text{area } FGDE}{\text{area } ABDE} = \frac{\log_e FG}{\log_e 10} = \log_{10} FG.$$

Huygens divides numerator and denominator by 32, which amounts to finding the 32nd root of each area, but this has the effect of so far closing up the figure that we may safely replace the hyperbola by a parabola whose outside area is known. He checks by finding a very good value for $\log_{10} 2$.

In the same year, 1661, Huygens finds another curve which he calls *logarithmic* but we should probably call it *exponential*. This curve has the property that the ordinate corresponding to the point midway between two given points of the x -axis is the mean proportional between their ordinates. The equation of the curve is $y = ka^x$. Huygens takes

$$y = 2^{x/x_0}; \quad x = \frac{\log y}{\log 2} x_0. \tag{6}$$

The constant subtangent is

$$\frac{ydx}{dy} = \frac{x_0}{\log_e 2}. \tag{7}$$

Huygens takes

$$x_0 = 10^n \log_{10} 2.$$

This gives for the constant subtangent

$$\log_{10} e = 0.43429448190325180,$$

“qualium logarithmus binarij est”

$$0.30102995663981195.$$

These numbers had long been known as they had appeared, for instance, in Briggs’ *Arithmetica logarithmica* of 1624, pages 10 and 14. As a matter

of fact, there appeared in 1618 a second edition of Wright’s translation of Napier’s *Mirifici Logarithmorum Canonis Descriptio* which contained an appendix, probably written by Oughtred, giving the natural logarithms of various numbers from 100,000 to 900,000. This is probably the earliest table of natural logarithms, although a very similar table by John Spidell appeared in 1619 [14].

The astonishing thing about all of those writers who connected logarithms with hyperbolic areas is their lack of interest in what we should call the base. Napier began by considering the relation between an arithmetical and a geometrical series. A geometrical series consists in successive powers of one number. What is that number? Or given a set of logarithms, what number has the logarithm 1? I mentioned that Briggs gave the logarithm of e , to the base 10 but I find no mention of e itself. Of course, we might write

$$\begin{aligned} 10^n \log_{10} \frac{10^n + \Delta x}{10^n} &= 10^n \log_e \left(1 + \frac{\Delta x}{10^n} \right) \log_{10} e \\ &= \Delta x \log_{10} e + \dots, \end{aligned}$$

but e itself does not appear. The fact is that there was no comprehension that a logarithm was essentially an exponent. Tropicke is very explicit in this point; he writes, “Freilich dürfen wir nicht an die moderne Erklärung der Logarithmen denken, die in ihnen Potenzexponenten einer bestimmten Grundzahl erkennt. Diese Auffassung machte sich erst um die Mitte des achtenten Jahrhunderts geltend” [15]. This is perhaps too strong a statement, for in a note on the same page he quotes James Gregory (whom he calls David Gregory) as saying in his *Exercitationes Geometricae* of 1684, p. 14, “Exponentes sunt ut logarithmi.” I have not been able to verify this, but we find in [16], “Si seriei Termonorum in Progressione geometrica ab 1 continue proportionalium, puta

$$1, 2, 4, 8, 16, 32, 64, \text{ etc.}$$

accomedetur series Indicum, sive Exponentium, in progressionem ab o continue procedentium, puta

$$0, 1, 2, 3, 4, 5, 6, \text{ etc.}$$

Hos exponentes appellabant Logarithmos.” We could not well ask for anything clearer or more explicit.

If most writers did not look on logarithms as exponents, how did they consider them? I think we find the clue in St. Vincent’s identification of logarithms with hyperbolic areas, remembering that these were

the days of Cavalieri and Roberval, when an area was looked upon as the same thing as an infinite number of line segments, a very helpful if dangerous definition. We find Halley writing [17], "They may more properly be said to be *numeri rationum exponentes*, wherein we consider ratio as a quantity *sui generis*, beginning from the ratio of equality, or 1 to 1 = 0, . . . and the *rationes* we suppose to be measured by the number of *ratiunculae* in each. Now these *ratiunculae* are in a continued scale of proportionals, infinite in number, between the two terms of the ratio, which infinite number of mean proportionals is to that infinite number of the like and equal *ratiunculae* between any other two terms as the logarithm of one ratio is to the logarithm of the other. Thus if we suppose there to be between 1 and 10 an infinite scale of mean proportionals whose number is 100000 *ad infinitum*, between 1 and 2 there shall be 30102 of said proportionals and between 1 and 3, 47712 of them which numbers therefore are the logarithms of the ratio of 1 to 10, 1 to 2, and 1 to 3, and so properly called the logarithms of 10, 2, and 3."

It is hard to see how there could be a much worse explanation of logarithms for those who "make constant use of logarithms without having an adequate notion of them." The one certain thing seems to be that a logarithm is an infinite number. I suppose we might translate this into the form

$$\frac{b}{a} = \frac{a + r_1}{a} \cdot \frac{a + r_2}{a + r_1} \cdot \frac{a + r_3}{a + r_2} \cdots \frac{a + r_n}{a + r_{n-1}} \cdot \frac{b}{a + r_n}$$

If

$$\frac{a + r_j}{a + r_{j-1}} = r, \quad \frac{b}{a} = r^n,$$

then n would be the logarithm.

4 Mercator, Newton, Leibniz

It is fair to say that such a definition of a logarithm was not original with Halley. We find Mercator writing in 1668, [18] "Est enim Logarithmus nihil aliud, quam numerus ratiuncularum contentarum in ratione quam absolutus quisque ad unitatem obtinet." I may mention also that this seems the first place where the words "logarithmus naturalis" are used. But the real significance of the article comes from the fact that instead of studying $\log x$ he takes up $\log(1 + x)$, which enables him to start from 0. The article is not clearly written, so I follow the much clearer exposition in Wallis [19], which was published in the same year.

We study the area under the curve (3) from $x = 1$ to $x = 1 + X$. We divide the length on the x -axis into n equal parts, each of length Δx . The abscissas are

$$1, 1 + \Delta x, 1 + 2\Delta x, \dots, 1 + X$$

and the corresponding ordinates are

$$1, \frac{1}{1 + \Delta x}, \frac{1}{1 + 2\Delta x}, \dots, \frac{1}{1 + (n - 1)\Delta x}.$$

The infinitesimal, rectangular areas are

$$\Delta x, \Delta x[1 - \Delta x + \Delta x^2 - \Delta x^3 + \dots],$$

$$\Delta x[1 - (2\Delta x) + (2\Delta x)^2 - (2\Delta x)^3 + \dots], \dots$$

Such infinite expansions were common in Wallis' work. The sum of these rectangular areas may be written

$$\begin{aligned} &\Delta x[1 + 1 + 1 + \dots] \\ &\quad - \Delta x[\Delta x + 2\Delta x + 3\Delta x + \dots] \\ &\quad + \Delta x[(\Delta x)^2 + (2\Delta x)^2 + (3\Delta x)^3 + \dots] - \dots \end{aligned}$$

Now $n\Delta x = X$, so we have

$$\begin{aligned} &X - \Delta x^2[1 + 2 + 3 + \dots] \\ &\quad + \Delta x^3[1^2 + 2^2 + 3^2 + \dots] - \dots \end{aligned} \quad (8)$$

With regard to these sums, Wallis says [19, page 222], "quod ostendit ille prop XVI etque a me alibi demonstratum." A reference he makes to Mercator is not conclusive as the statement is sketchy; as to his own work I will follow [20], as I shall need that again. Here he is seeking the area under the curve $y = x^n$ from $x = 0$ to $x = X$. His method is not perfectly clear, as he seems merely to generalize by analogy from cases worked out earlier, but what he does is essentially the following:

We take N equal lengths from 0 to $N\Delta x = X$. We have a set of rectangles whose combined areas are

$$\Delta x[0^m + (\Delta x)^m + (2\Delta x)^m + (3\Delta x)^m + \dots].$$

Let us assume that

$$\begin{aligned} &0^m + 1^m + \dots + (N - 1)^m \\ &\quad = \alpha N^{m+1} + \beta N^m + \gamma N^{m-1} + \dots \end{aligned}$$

Replacing N by $N + 1$, and subtracting, we obtain

$$N^m = (m + 1)\alpha N^m + \beta N^{m-1} + \gamma N^{m-2} \dots,$$

so

$$\alpha = \frac{1}{m + 1}.$$

Substituting, and remembering that $N\Delta x = X$, there results

$$\text{area} = \frac{X^{m+1}}{m + 1} + \beta\Delta x X^m + \gamma\Delta x^2 X^{m-1}.$$

The limit of this as $N \rightarrow \infty$ is $X^{m+1}/m + 1$, since $\Delta x \rightarrow 0$. We thus can substitute this result in (8), when $m = 1, 2, 3, \dots$, to obtain Mercator's famous formula:

$$\log(1 + X) = X - \frac{X^2}{2} + \frac{X^3}{3} - \frac{X^4}{4} + \dots \quad (9)$$

A good deal has been written about this series, as we see from Mazerés and elsewhere. The obvious way to obtain the equation is to apply the calculus, so we now turn to see how this instrument was brought to bear. In 1669, a year after Mercator had published his work on logarithms [18], Newton sent to Collins his article, *De Analysi per aequationes numero terminorum Infinitas* [21]. This represents his first studies of areas under curves, which he had been working at for a year or two, but had not published. In fact, publication did not occur for a goodly number of years to come; there is, however, no question of giving his results precedence over those of Mercator. It begins as shown below:

Curvarum simplicium Quadratura

Reg. 1: Si $a^{m/n} = y$, erit $\frac{an}{m+n} x^{(m+n)/n} = \text{area } ABD$.

I must speak further of this. In [22] we read on p. 176, "Dr. Wallis published in his *Arithmetica infinitorum* in the year 1655 and in the 59th Proposition of that Book, if the Abscissa of any curvilinear figure be called x and m and n be two Numbers, the ordinates erected at right Angles be $x^{m/n}$ the area of the Figure shall be $(n/(m+n))x^{(m+n)/n}$. And this is assumed by Mr. Newton, upon which he finds his Quadrature of Curves. Dr. Wallis demonstrated this by steps in many particular Propositions and then connected all the Propositions into one by a Table of Cases. Mr. Newton reduced all Cases to One, with an indefinite Index, and at the end of his Compendium demonstrated it at once by his method of moments, he being the first who introduced indefinite Indices of Dignities into the Operations of Analysis." This is Newton's own statement of the case and must be taken as final. It is true that Wallis

worked out a number of special cases in a manner not exactly like the method followed here, and did not use a literal exponent. The greater generality of Newton's formula is found by replacing x by $x^{1/n}$. Newton's proof by "the method of moments" we should call differentiation, and consisted in showing that

$$\text{if } z = \frac{n}{m+n} x^{(m+n)/n}, \text{ then } \frac{dz}{dx} = x^{m/n}.$$

It is fair to say also that although he gives Mercator's formula, he gives it as the area under the hyperbola, with no mention of Mercator or of logarithms.

It is time to turn for a moment to the other inventor of the calculus, Gottfried Leibniz. We find him writing in 1677 or 1678 [23], "In Hyperbol sit

$$AB = 1, BM = x, ML = \frac{1}{1+x},$$

$$CBMLC = \frac{1}{1}x - \frac{1}{2}x^2 + \frac{1}{3}x^3 - \frac{1}{4}x^4 + \dots"$$

This is proved by the straight expansion of $1/(1+x)$, after which there is integration term by term. We find something more interesting a dozen years later, when he writes to Huygens, who is said never to have understood Leibniz's calculus of differences [24], "Soit donc x l'abscisse et y l'ordonnée de la courbe, et l'équation comme je vous ay dit

$$\frac{x^3y}{h} = b^{2xy}.$$

Je désignerai le logarithme de x par $\log x$ et nous aurons

$$3 \log x + \log y - \log h = 2xy$$

supposant que le log de l'unité soit 0 et le log $b = 1$. Donc par la quadrature de l'hyperbole nous aurons

$$3 \int \frac{dx}{x} + \int \frac{dy}{y} - \log h = 2xy$$

$$3 \frac{dx}{x} + \frac{dy}{y} = 2xdy + 2ydx,$$

dx sera à dy , on bien DB sera à y comme $2x^2y - x$ est à $3y - 2xy^2$ c'est à dire DB sera

$$\frac{2x^2y - xa^2}{3a^2 - 2xy}$$

comme vous le demandées, a estant l'unité."

5 Leonhard Euler

It is now time to turn to the man who pulled all this together and who put the number e definitely on the map, Leonhard Euler. This he did in [25], beginning in “Caput VII” with the base a . His argument is outlined below:

Since $a^0 = 1$, we may put

$$a^w = 1 + kw; \quad w = \log(1 + kw).$$

Assume w to be very small, and write

$$\begin{aligned} a^{iw} &= (1 + kw)^i \\ &= 1 + \frac{i}{1}kw + \frac{i(i-1)}{1 \cdot 2}k^2w^2 \\ &\quad + \frac{i(i-1)(i-2)}{1 \cdot 2 \cdot 3}k^3w^3 + \dots \end{aligned}$$

Since w is infinitesimally small, and i is infinitely large, we write $iw = z$

$$\begin{aligned} a^z &= \left(1 - \frac{kz}{i}\right)^i \\ &= 1 + kz + \frac{(i-1)}{i \cdot 1 \cdot 2}k^2z^2 + \frac{(i-1)(i-2)}{i \cdot 1 \cdot 2 \cdot 3}k^3z^3 + \dots \end{aligned}$$

Since i is very large, we may assume $(i-n)/i = 1$, then

$$a^z = 1 + kz + \frac{k^2z^2}{1 \cdot 2} + \frac{k^3z^3}{1 \cdot 2 \cdot 3} + \dots$$

If $z = 1$,

$$a = 1 + k + \frac{k^2}{1 \cdot 2} + \frac{k^3}{1 \cdot 2 \cdot 3} + \dots$$

If we take $a = 10$, the base in the logarithm system of Briggs, Euler gives $k = 2.30238$, approximately.

For a natural logarithm we take $k = 1$; $a = e$; and

$$e = 1 + \frac{1}{1} + \frac{1}{1 \cdot 2} + \frac{1}{1 \cdot 2 \cdot 3} + \dots \quad (10)$$

Euler gives this value to 18 places, without naming the source, namely,

$$e = \lim_{i \rightarrow \infty} \left(1 + \frac{1}{i}\right)^i \quad (11)$$

With regard to the use of the letter e , Euler had long employed it, for we find him writing [26], page 80, “scribitur pro numero cujus logarithmus est unitas e , qui est 2.7182817...” Note that this is Leibniz’s b .

I pass to Ch. VII of the *Introductio*. Euler assumes for small values of z ,

$$\sin z = z, \quad \cos z = 1.$$

He then, following DeMoivre, writes,

$$\begin{aligned} \cos nz &= \frac{(\cos z + \sqrt{-1} \sin z)^n + (\cos z - \sqrt{-1} \sin z)^n}{2}, \\ \sin nz &= \frac{(\cos z + \sqrt{-1} \sin z)^n - (\cos z - \sqrt{-1} \sin z)^n}{2\sqrt{-1}}. \end{aligned}$$

Putting $nz = v$, and remembering that z is small,

$$\begin{aligned} \cos v &= 1 - \frac{v^2}{1 \cdot 2} + \frac{v^4}{1 \cdot 2 \cdot 3 \cdot 4} - \dots, \\ \sin v &= v - \frac{v^3}{1 \cdot 2 \cdot 3} + \frac{v^5}{1 \cdot 2 \cdot 3 \cdot 4 \cdot 5} - \dots, \end{aligned}$$

Comparing these with the value given previously for a^z , one obtains

$$\cos v = \frac{e^{v\sqrt{-1}} + e^{-v\sqrt{-1}}}{2}; \quad (12)$$

$$\sin v = \frac{e^{v\sqrt{-1}} - e^{-v\sqrt{-1}}}{2\sqrt{-1}}; \quad (13)$$

and

$$v\sqrt{-1} = \log[\cos v + \sqrt{-1} \sin v]. \quad (14)$$

This last formula was not, strictly speaking, original. Roger Cotes in [27] sought the area of an ellipsoid of revolution. When the rotation is about the minor axis there is no trouble, but when the motion is about the major axis we find him writing “Posset hujus etiam superficiei per Logometriam designari, sed modo inexplicabili... arcus erit rationis inter

$$EX + XC\sqrt{-1}aCE \text{ mensura ducta in } \sqrt{-1}.”$$

I will leave Euler for a moment to speak of the numerical value of e . William Shanks, who, until quite recently, held the world’s record of 707 places for π , had a try at e [28]. Glaisher found an error in this, but Shanks corrected it, and calculated a value which he was sure was right to 205 places. Glaisher verified 137 of them. Boorman [29] calculated e to 346 places. He acknowledged that he and Shanks agreed only up to 187 places. “One is wrong, which one?” Boorman gives the impression of being a rather amateurish mathematician. Adams [30] calculated $\log_{10} e$ to 272 places, probably all

correct. Many years ago I knew a youthful teacher of mathematics who had the vaulting ambition to calculate e by long-hand methods to 1,000 places. I lost sight of him over fifty years ago, probably he died early of heart failure.

I return to Euler. In Caput XVIII on *De Fractionibus continuis* [25], he describes methods of expansion into a continued fraction. When it is a question of turning a rational fraction into a continued one, the process is essentially that of finding a highest common factor, and can be done in only one way. Euler writes

$$e = 2.718281828459 \dots,$$

$$\frac{e-1}{2} = -0.8591409142295.$$

He writes this in the form,

$$\frac{e-1}{2} = \frac{1}{1 + \frac{1}{6 + \frac{1}{10 + \frac{1}{14 + \frac{1}{18 + \frac{1}{\text{etc.}}}}}}}}$$

and remarks [25], page 388, “Cuius fractio ex Calculo infinitesimali dari potest.”

Euler assumes that the quotients will increase by 4 each time, so that the fraction goes on indefinitely. Hence e is not a rational fraction.

As for finding this “ex Calculo infinitesimali” he returns to this very much later in life, “Summatio fractionis continuæ cujus indices progressionem arithmeticam constituunt” in Vol. 23 of his *Opera* mentioned in [25]. The method consists in establishing contact with a Riccati differential equation. For a fuller discussion see [31]. Euler did not complete all the details with modern rigor, but what I have just shown is the first attempt to demonstrate the irrationality of e .

We must wait a whole century for anything really new and startling in this line. This came in 1874 with Hermite’s proof that e is not an algebraic number [32], that is, not the root of any equation with integral coefficients. A much simpler demonstration is given by Klein in [33].

References

1. Archimedes, *Opera omnia*, 3rd ed. Heiberg, vol. III, 1915, p. 79.
2. G. J. Allman, *Greek Geometry*, Dublin, 1889.
3. G. Loria, *Le Scienze esatte nella antica Grecia*, 2nd ed. Milan, 1914, p. 155
4. T. L. Heath, *The Works of Archimedes*, Cambridge, 1897, p. 1.
5. H. Zeuthen, *Die Lehre von den Kegelschnitten*, Kopenhagen, 1886, p. 463.
6. Die Kegelschnitte des Gregorius a St. Vincento, *Abhandlungen zur Geschichte der mathematische Wissenschaften*, Vol. XIX, Part 2, Leipzig, 1907.
7. Abraham Gotthilf Kastner, *Geschichte der Mathematik*, Vol. 3, Göttingen, 1799.
8. G. Cantor, *Geschichte der Mathematik*, Vol. 2, 2nd Ed., Leipzig, 1900, p. 715.
9. Charles Hutton, *Mathematical Tables*, London, 1804, p. 80.
10. J. Wallis, *Opera Mathematica*, Oxford, 1693, Vol. 2.
11. *Philosophical Transactions*, Abridged, Vol. I, p. 232.
12. C. Montucla, *Histoire des mathématiques*, Vol. 2, Paris, 1800, p. 80.
13. C. Huygens, *Oeuvres Complètes*, Vol. 14, La Haye, 1920, pp. 433, 441, and 474.
14. Glaisher, The earliest use of the Radix Method for calculating logarithms, *Quarterly Journal of Mathematics*, Vol. 46, 1914–15, especially p. 174.
15. H. Tropicke, *Geschichte der Elementarmathematik*, 3rd Ed., Vol. 2, Berlin, 1933, p. 205.
16. J. Wallis, *Algebra*, Ch. XII, *Opera*, Vol. 2, Oxford, 1693, pp. 57, 58.
17. A most compendius and facile Method for constructing Logarithms exemplified and demonstrated from the Nature of Numbers, *Philosophical Transactions*, abridged, Vol. IV. 1695-1702, London, 1809, p. 19.
18. *Logarithmo-technica Auctore Nicolao Mercatore*; see Mazerés, *Scriptores Logarithmici*, Vol. 1, London, 1791, p. 169.
19. J. Wallis, Logarithmo-technica Nicola Mercatoris, *Philosophical Transactions*, August, 1668. Mazerés *cit.* in [18], p. 221.
20. —, *Arithmetica Infinitorum*, Oxford, 1656. Especially Prop. 59.
21. I. Newton, *Commercium Epistolicum Collinsii et aliorum*, published by Biot and Leffort, Paris, 1856.
22. An Account of the Book entitled *Commercium Epistolicum Collinsii et aliorum*, Anonymously by Newton, *Philosophical Transactions*, Vol. XXIX, London.
23. *Leibnizens Mathematische Schriften*, Gerhardt Ed., Part 2, Vol. 1. Halle, 1858.
24. *Ibid.*, Part I, Vol. 2.

25. L. Euler, *Introductio in Analysin infinitorum*, Lausanne, 1748, also his *Opera Omnia seria prima, Opera mathematica*, Vol. 8.
26. —, *Meditatio in Experimenta explosione Opera Postuma*, Petropoli, 1862.
27. R. Cotes, *Harmonia Mensurarum*, Cambridge, 1722, p. 28.
28. *Proceedings of the Royal Society*, Vol. 6, 1854.
29. Computation of the Napierian Base, *Mathematical Magazine*, Vol. 1, 1884, p. 204.
30. Shanks, On the Modulus of Common Logarithms, *Proceedings of the Royal Society*, Vol. 43, 1887.
31. Pringsheim, Ueber die ersten Beweise der Irrationalität von e und π , *Sitzungsberichte der K Akademie der Wissenschaften zu München*, Vol. 28, 1898.
32. Charles Hermite, *Sur la fonction exponentielle*, Paris, 1874; *Oeuvres*, Vol. III, Paris, 1912.
33. Klein, *Ausgewählte Fragen der Elementargeometrie*, Leipzig, 1895, pp. 47ff.

Euler's Vision of a General Partial Differential Calculus for a Generalized Kind of Function

JESPER LÜTZEN

Mathematics Magazine 56 (1983), 299–306

The vibrating string controversy involved most of the analysts of the latter half of the 18th century. The dispute concerned the type of functions which could be allowed in analysis, particularly in the new partial differential calculus. Leonhard Euler held the bold opinion that all functions describing any curve, however irregular, ought to be admitted in analysis. He often stressed the importance of such an extended calculus, but did almost nothing to support his point of view mathematically. After having been abandoned during the introduction of rigor in the latter part of the 19th century, Euler's ideas began to take more concrete form during the early part of the 20th century, and they have now been incorporated into L. Schwartz's theory of distributions.

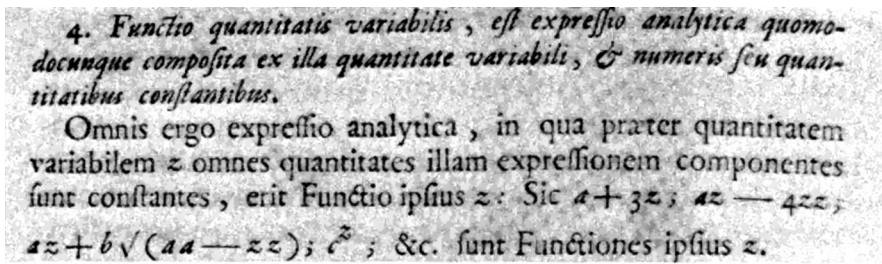
1 The algebraic function concept

Euler's radical stand in the dispute over the vibrating string is surprising since he had canonized the narrower range of analysis which his main opponent, J. B. R. d'Alembert (1717–1783), adhered to. This was done in the influential book *Introductio in analysin infinitorum* [12], in which Euler chose to determine the relation between the variable quantities by way of functions instead of using curves, as had been universally done earlier (cf. [22] and [7]). He defined a function as follows (see photo on the next page):

A function of a variable quantity is an analytical expression composed in one way or another of this variable quantity and numbers or constant quantities [12, ch. 1, §4].

In forming the analytical expressions, Euler allowed the use of the standard transcendental operations such as log, exp, sin and cos in addition to algebraic operations. Still, all the rules in the theory of functions were taken over from algebra, so that Euler's function concept was in essence entirely algebraic. Thus *Introductio* marked a shift in the setting of analysis from geometry to algebra. Euler even accepted, and treated algebraically, infinite expressions





such as infinite series, infinite products and continued fractions. Lebesgue [25] later showed that when such infinite limit procedures are accepted, the class of functions is very extensive, namely, equal to the class of Borel Functions. However, Euler did not realize the immense generality of his function concept and in theoretical considerations he conferred on them all the nice properties he needed such as differentiability and even analyticity in the modern sense. Still, it would be off the mark to identify Euler's functions with one of the modern classes of functions such as differentiable functions or analytical functions because their definition involves topological (geometrical) ideas which are foreign to Euler's way of thinking.

Most important among the nice properties shared by all Euler's functions was the possibility of expanding them in a power series:

$$f(x + i) = f(x) + pi + qi^2 + ri^3 + \dots,$$

for in all differentiations actually carried out in Euler's second influential textbook *Institutiones calculi differentialis* [16] the differential quotient is found as the coefficient p of the first power term. Later in the century J. L. Lagrange [24] defined the derivative of a function in this way and gave a "proof" that the expansion always exists. In the mid-18th century, however, power series were only used as a practical tool whereas the metaphysical basis for the calculus was found elsewhere. For example, d'Alembert defined the derivative using limits, and Euler's definition of the differential rested on a theory of zeros of different order. Yet, these foundational differences were not reflected in the domain they assigned to the ordinary calculus; both agreed that

... [calculus] as it has been treated until now can only be applied to curves, whose nature can be contained in one analytical equation [18, §7].

2 Euler's generalized functions

The discussion of the vibrating string brought an end to this agreement. D'Alembert, who in 1747 [1] found his famous solution

$$y = f(x, t) = \phi(x + t) + \psi(x - t)$$

of the wave equation

$$\frac{\partial^2 f}{\partial x^2} = \frac{\partial^2 f}{\partial t^2}$$

governing the displacement y of the string, required that the "arbitrary" functions ϕ and ψ be analytical expressions.

In all other cases the problem cannot be solved, at least not with my method, and I do not even know whether it will not be beyond the powers of the known analysis. In fact, it seems to me that one cannot express y analytically in a more general way than supposing it to be a function of x and t [2, p. 358].

Euler, on the other hand, pointed out that this requirement restricted the initial displacement $\phi(x) + \psi(x)$ of the string too much; for example, he believed that the plucked string (Figure 1) would be excluded from d'Alembert's solution. (However, the plucked string can be described analytically by a slight modification of Cauchy's example: $\sqrt{x^2} = |x|$ [9].) Therefore he argued that one had to allow the functions ϕ and ψ to represent arbitrarily given curves. In this way physical reality led Euler to generalize the function concept so as to be in one to one correspondence with the geometrical concept of

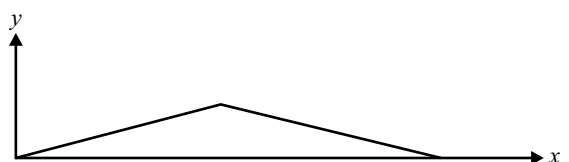


Figure 1.

curve which he had earlier abandoned as the basic concept in analysis.

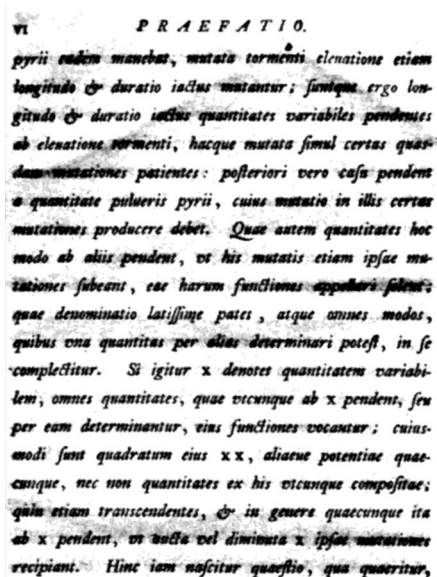
It is surprising that Euler never provided a proper definition of the more general notion of function. His many papers on the vibrating string (particularly [17]) made clear that a generalized function was something corresponding to a general hand-drawn curve, but he never explicitly stated what this something was supposed to be. To judge from the classification of the new functions he seems to have had an algebraic definition in mind. He divided the general functions into the continuous and the discontinuous. The former were identical with the functions defined in *Introductio*, whereas the latter could not be expressed by one analytical expression. Euler was quite explicit about the continuity of a function having nothing to do with the connectedness of the curve; for example $1/x$ is continuous but its graph is disconnected at $x = 0$. Thus Euler's concept of continuity must be distinguished from the modern concept, due to Cauchy [8], and so we shall term the former ***E*-continuity**. In [17] Euler further divided the *E*-discontinuous functions into mixed functions, whose graph can be represented piecewise by finitely many analytical expressions, and the functions corresponding to arbitrary hand-drawn curves, whose analytical expressions may, so to speak, change from point to point.

Thus Euler's division of functions into classes was entirely algebraic and so was his distinction between even and odd functions. For example, in his critique [15] of D. Bernoulli's [6] description of the vibrating string as a trigonometric series, Euler argued that an *E*-discontinuous function of the form

$$\begin{cases} f(x) & \text{for } x > 0 \\ -f(-x) & \text{for } x < 0 \end{cases}$$

is only odd if f is odd and by that he meant that its power series contains only odd powers of x . To conclude: even when the consequences were absurd, Euler continued to think algebraically about his new functions, which, implicitly, he defined as the collection of the (possibly infinitely many) analytical expressions describing the corresponding curve.

Strangely enough, Euler himself had introduced a way of thinking about functions which he could have used to define his *E*-discontinuous functions as separate entities. In his second textbook on analysis *Institutiones calculi differentialis* (1755) [16], he defined functions in the following way (see photo above):



If, therefore, x denotes a variable quantity, all quantities which depend in some way on x or are determined by it, are called functions of this variable [16, Preface].

As it stands, this is almost the modern function definition and it clearly encompasses the *E*-discontinuous functions. However, Euler did not realize its generality. In *Institutiones calculi differentialis* only *E*-continuous functions occur, and the *E*-discontinuous functions are not even mentioned. Neither did he refer to his 1755 definition in any of his later papers on *E*-discontinuous functions. This indicates that Euler thought of his 1755 function definition as being equivalent to the definition given in *Introductio*. In fact, Euler's statement from 1765 (quoted earlier) that analysis until then had exclusively been concerned with analytical expressions only makes sense under this assumption. (This point of view is different from the one put forward by Youschkevich [34].)

3 Euler's vision of a generalized calculus

The lack of a proper definition of the *E*-discontinuous functions suggests that Euler's main concern was not the foundation of the generalized function concept itself but the analysis it made possible. We saw that initially Euler had introduced his new functions for physical reasons. Later [17] he stressed that the *E*-discontinuous functions were

not forced onto analysis from outside but inevitably emerged as arbitrary functions in the partial integral calculus. For example [20, book 2, sect. 1, §33], the solution of the partial differential equation

$$\frac{\partial u(x, y)}{\partial x} = 0$$

is an arbitrary constant under the variation of x , but the constant can vary as a function f of y . It does not matter whether the constants for different values of y are connected by an analytical expression or not; therefore f must be allowed to be E -discontinuous. Since the functions ϕ and ψ in the solution of the wave equation arise in this way when $x+t$ and $x-t$ are used as independent variables, these functions are by their nature general functions.

Euler only used the E -discontinuous functions in the calculus of functions of several variables, but within that theory he would apparently blaze the trail for their unrestricted application. In contrast to the conservative d'Alembert, Euler argued that the development of a calculus of E -discontinuous functions is particularly desirable because all earlier calculus had been restricted to analytic expressions:

But if the theory [of the vibrating string] leads us to a solution so general that it extends to all discontinuous as well as continuous figures, one must admit that this research opens to us a new road in analysis by enabling us to apply the calculus to curves which are not subject to any law of continuity, and if that has appeared impossible until now the discovery is so much more important [18, §8].

Euler's insistence that calculus should be applicable within the whole new function domain instead of being restricted to some—possibly varying—subclass(es) (as is the case in modern analysis) was supported not only by the mentioned physical reasons. It was also in agreement with the fundamental belief in the generality of mathematics. For algebraic rules were considered universally valid because they operated on abstract quantities, and since analysis was just infinite algebra, its rules had to be generally applicable as well.

For, because this calculus applies to variable quantities, that is, quantities considered generally, if it were not generally true. . . one could never make use of this rule, since the truth of the differential calculus is based on the generality of the rules of which it consists [14, 1. Objection].

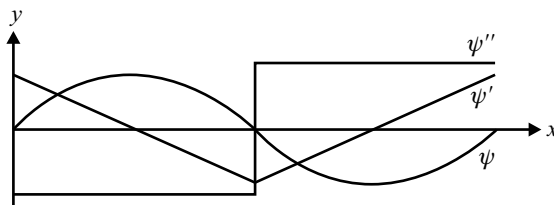


Figure 2.

This basic belief in the generality of mathematics forced Euler to extend calculus to all E -discontinuous functions as soon as he had allowed them to enter his mathematical universe. Initially it probably also made him believe that this extension would come down to a simple admission of all the well-known rules to the extended domain. However, he soon had to realize that d'Alembert's exclusion of E -discontinuous functions was not only due to plain conservatism but was supported by mathematical arguments.

In many examples d'Alembert showed that the mathematical analysis of the vibrating string broke down at points where ϕ or ψ changed their analytical expression. For example, d'Alembert [3, §7] proved that if ψ is composed of two symmetric parabolas as in Figure 2 and $\phi \equiv 0$ then $\psi(x-t)$ does not satisfy the wave equation

$$\frac{\partial^2 f}{\partial x^2} = \frac{\partial^2 f}{\partial t^2}$$

at points where $x-t=0$. This and other difficulties can be explained in modern terminology by the fact that ϕ or ψ are not twice differentiable. D'Alembert came close to such an insight towards the end of his life [4], but while the controversy was at its highest, he believed that he had proved that ϕ and ψ must be E -continuous.

Euler was not convinced by d'Alembert's arguments and tried to refute them with a few counterarguments [19] of which I shall reproduce the most convincing. He remarked that the trouble was due to the sharp bend in the first derivative of ψ . Therefore, one had only to smooth out ψ' which could be done by changing ψ infinitely little to $\tilde{\psi}$. Since $\tilde{\psi}(x-t)$ would then satisfy the wave equation, one also had to admit $\psi(x-t)$ as a solution since infinitely small changes were always ignored in analysis.

In Eulerian calculus this argument is not completely off the mark, and even in modern analysis it contains the germ of a good idea (cf. following page). Still Euler seems to have realized that he had not overcome all objections to his new general anal-

ysis, and so he often encouraged the younger mathematicians to work on these problems.

This part of analysis [of two or more variables] is essentially different from the former [of one variable], and extends even to functions void of all law of continuity. This part, of which we so far know barely the first elements, certainly deserves the united efforts of all geometers for its investigation and development [19, §32].

4 The fate of Euler's vision

In order to follow how subsequent geometers cultivated this new branch of analysis it is useful to divide the complex of problems, seen by Euler as a unity, into three separate parts:

- (1) The generalization of the concept of function.
- (2) The generalization of analysis.
- (3) The development of the theory of partial differential equations.

The last and most important point of this research programme (3) was enthusiastically taken up by most of the mathematical community and was probably the most important mathematical discipline during the following half century. However, a discussion of it is far beyond the scope of this paper (see [23, ch. 22, 28]).

The generalization of the function concept (1) was also gradually accepted. In this process Euler's 1755 function definition was influential, regardless of his own interpretation of it. For after 1755 it became normal to reproduce this definition in textbooks on analysis, and slowly mathematicians began to realize its true generality. But this process took almost a century. For example, Lagrange [24] and Cauchy [8] defined functions generally as correspondences between variables, but they both thought of them as analytical expressions. It is natural in Lagrange's case, because he carried Euler's algebraic approach to its extreme, but it is surprising that the father of modern analysis, Cauchy, had a similar way of thinking. Still, this is evident from many remarks in his famous *Cours d'Analyse* [8], for example, the talk about "the constants or variables contained in a given function" [8, ch. 8, §1].

In J. Fourier's works [21, §417], one can find some comprehension of the generality of Euler's 1755 definition but the first mathematician who really took it seriously and understood the implications of the permissible pathologies was J. P. G. Lejeune-

Dirichlet [11], after whom our function concept is justly named.

The generalization of analysis (2) suffered the opposite fate. At first it gained widespread acceptance but during the 19th century the idea was entirely abandoned. It happened as follows. In 1787 the St. Petersburg Academy officially terminated the controversy over the vibrating string by awarding L. Arbogast the first prize for a paper on the irregularities of arbitrary functions in the solutions of partial differential equations. Arbogast came out in favor of Euler's point of view, but he added nothing new to the foundational difficulties [5].

However, this official support of a general calculus was brushed aside by Cauchy, whose partial rigorization of analysis was a frontal attack on the principle of the generality of algebraic and analytical rules which had philosophically supported Euler's point of view. Cauchy explicitly pointed out this fundamental shift in the introduction to his famous *Cours d'Analyse* [8]:

As for the methods, I have tried to give them all the rigour that one demands in geometry, so as never to have recourse to reasoning drawn from the generality of algebra.

Therefore nothing in his philosophy prevented him from confining calculus to a subclass of the class of functions, and in essence he restricted its use to the continuous functions (in the modern sense). In some of his papers he realized the inadequacy of this restriction, but a clear idea of the spaces $C^n(\mathbf{R})$ as the domain of d^n/dx^n did not crystalize until the 1870s in the Weierstrass school.

As a whole, mathematics benefited from this rigorization of analysis, but the corresponding restriction in the allowable solutions to partial differential equations made life complicated for the applied mathematician. Thus when irregular physical situations occurred (as, for example, a sharp bend in a string), the differential equation could not be used and a new mathematical model of the system had to be found. Such alternative models were set up, for example, by E. Christoffel [10].

However, in the beginning of the 20th century this procedure was felt to be so cumbersome and unnatural that several definitions of generalized solutions to partial differential equations were suggested, beginning in 1899 with H. Petrini's generalization of Poisson's equation [28]. Of the many generalization procedures I shall mention only the "sequence definition" implicitly used by N. Wiener in 1926 [33]

and explicitly introduced by Sobolev (1935) [32]. According to this definition, f is a generalized solution to a (partial) differential equation if there exists a sequence of ordinary solutions $\{f_n\}$ converging, in a suitable topology, to f . This definition is particularly interesting because it leads to a sensible interpretation of Euler's argument against d'Alembert; for, if instead of one smooth function $\tilde{\psi}$ infinitely close to ψ , we think of a sequence ψ_n of such functions, then Euler's argument shows that $\psi(x-t)$ is a generalized solution to the wave equation.

All the ad hoc definitions of generalized solutions from the first half of this century were incorporated in the theory of distributions created by L. Schwartz during the period 1945-1950 [31] as a result of his work with generalized solutions to the polyharmonic equation [30]. The theory of distributions probably constitutes the closest approximation to Euler's vision of a general calculus one can obtain, for in that theory any generalized function is infinitely often differentiable. However, in many respects the reality has turned out to be different from the dream. In one respect the reality is more satisfactory since it not only generalizes partial differential calculus which Euler had imagined but encompasses ordinary differential calculus as well. In other respects it is less perfect; for example, the general use of the algebraic operations, such as multiplication of two generalized functions, has been sacrificed in the theory of distributions. Moreover, the necessary generalization of the function concept has turned out to be much more extensive than the one Euler suggested.

5 Concluding remarks

Surely the realization of Euler's vision of a general calculus was different from what he had imagined—and more difficult. This can only increase our admiration for his readiness to overthrow his own framework of analysis when physical reality called for it. His conduct reveals an undogmatic and flexible attitude toward the foundational problems, from which much could be learned by modern mathematicians. On the other hand, it is worth noting that the creation of the theory of distributions made extensive use of the classical theory of differential operators created more in the spirit of d'Alembert; one can even argue that the establishment of a secure foundation for the more restricted classical calculus was a necessary condition for the realization of Euler's vision of a general calculus.

As further reading on the development of the concept of function I can recommend [34], [29] and, for those who want to brush up their Danish, [26]. The book [27] contains more information on the history of generalized solutions to partial differential equations and other aspects of the prehistory of the theory of distributions.

References

1. J. d'Alembert, Recherches sur la courbe que forme une corde tendue mise en vibration, *Mém. Acad. Sci. Berlin*, 3 (1747) 214–219.
2. —, Addition au mémoire sur la courbe que forme une corde tendue mise en vibration, *Mém. Acad. Sci. Berlin*, 6 (1750) 355–366.
3. —, Recherches sur les vibrations des cordes sonores, *Opuscules Mathématiques*, 1 (1761) 1–73.
4. —, Sur les fonctions discontinues, *Opuscules Mathématiques*, 8 (1780) 302–308.
5. L. F. A. Arbogast, *Mémoire sur la nature de fonctions arbitraires qui entrent dans les intégrales des équations aux différences partielles*, St. Petersburg, 1791.
6. D. Bernoulli, Réflexions et éclaircissemens sur les nouvelles vibrations des cordes, *Mém. Acad. Sci. Berlin*, 9 (1753 publ. 1755) 147–172 (see also 173–195).
7. H. J. M. Bos, Differentials, higher-order differentials and the derivative in the Leibnizian calculus, *Arch. Hist. Exact Sci.*, 14 (1974) 1–90.
8. A.-L. Cauchy, *Cours d'analyse de l'école roy. Polytechnique*, 1re partie; Analyse algébrique, Paris, 1821 = *Oeuvres* (2) 3.
9. —, Mémoire sur les fonctions continues ou discontinues, *Comp. Rend. Acad. Roy. Sci. Paris*, 18 (1844) 145–160 = *Oeuvres* (1) 8, 145–160.
10. E. Christoffel, Untersuchungen über die mit Fortbestehen linearer partieller Differentialgleichungen verträglichen Unstetigkeiten, *Ann. Mat. Pur. Appl.*, (2) 8 (1876) 81–112 = *Gesammelte Math. Abh.*, 2, 51–80.
11. J. P. G. Lejeune Dirichlet, Sur la convergence des séries trigonométriques qui servent à représenter une fonction arbitraire entre les limites données, *J. Reine Angew. Math.*, 4 (1829) 157–169 = *Werke* I, 117–132.
12. L. Euler, *Introductio in analysin infinitorum* (2 vols), Lausanne, 1748 = *Opera Omnia* (1) 8, 9.
13. —, Sur la vibration des cordes, *Mém. Acad. Sci. Berlin*, 4 (1748, publ. 1750) 69–85 = *Opera Omnia* (2) 10, 63–77.

14. —, De la controverse entre Messieurs Leibniz et Bernoulli sur les logarithmes des nombres négatifs et imaginaires, *Mém. Acad. Sci. Berlin*, 5 (1749) 139–179 = *Opera Omnia* (1) 17, 195–232.
15. —, Remarques sur les mémoires précédens de M. Bernoulli, *Mém. Acad. Sci. Berlin*, 9 (1753, publ. 1755) 196–222 = *Opera Omnia* (2) 10, 233–254.
16. —, *Institutiones calculi differentialis*, St. Petersburg, 1755 = *Opera Omnia* (1) 10.
17. —, De usu functionum discontinuarum in analysi, *Nov. Comm. Acad. Sci. Petrogr.*, 11 (1763, publ. 1768) 67–102 = *Opera Omnia* (1) 23, 74–91.
18. —, Eclaircissemens sur le mouvement des cordes vibrantes, *Miscellanea Tourinensia*, 3 (1762–1765 publ. 1766) math. cl., 1–26 = *Opera Omnia* (2) 10, 377–396.
19. —, Sur le mouvement d'une corde qui au commencement n'a été ébranlée que dans une partie, *Mém. Acad. Sci. Berlin*, 21 (1765 publ. 1767) 307–334 = *Opera Omnia* (2) 10, 426–450.
20. —, *Institutiones calculi integralis* (3 vols), St. Petersburg, 1768–1770.
21. J. B. J. Fourier, *Théorie Analytique de la Chaleur*, Paris, 1822 = *Oeuvres* I.
22. G. F. A. Hospital, *Analyse des Infiniments Petits pour l'intelligence des lignes courbes*, Paris, 1696.
23. M. Kline, *Mathematical Thought from Ancient to Modern Times*, Oxford: Oxford Univ. Press, 1972.
24. J. L. Lagrange, *Théorie des Fonctions Analytiques*, Paris, 1797, 2nd ed. 1813 = *Oeuvres* 9.
25. H. Lebesgue, Sur les fonctions représentables analytiquement, *J. Math. Pures Appl.*, 1 (1905) 139–216.
26. J. Lützen, Funktionsbegrébets udvikling fra Euler til Dirichlet, *Nordisk Mat. Tidsskr.*, 25–26 (1978) 5–32.
27. —, *The Prehistory of the Theory of Distributions*, Berline: Springer, 1982.
28. H. Petrini, “Démonstration générale de l'équation de Poisson $\Delta V = -4\pi\rho$ en ne supposant que ρ soit continu,” *K. Vet. Akad. Oeuvres*, Stockholm, 1899.
29. J. R. Ravetz, Vibrating strings and arbitrary functions, *Logic of personal knowledge: Essays presented to M. Polanyi on his 70th birthday*, London, 1961, 71–88.
30. L. Schwartz, Sur certaines familles non fondamentales de fonctions continues, *Bull. Sec. Math. France*, 72 (1944), 141–145.
31. —, *Theorie des Distributions* (2 vols), Paris: Hermann, 1950, 1951.
32. S. L. Sobolev, Obshchaya teoriya difraktsii voln na rimanovykh poverkhnostyakh, *Travaux Inst. Steklov. Tr. Fiz.-Mat. in-ta*, 9 (1935) 433–438.
33. N. Wiener, The operational calculus, *Math. Ann.*, 95 (1926) 557–585.
34. A. P. Youshkevich, The concept of function up to the middle of the 19th century, *Arch. Hist. Exact Sci.*, 16 (1976) 37–85.

Euler and the Fundamental Theorem of Algebra

WILLIAM DUNHAM

College Mathematics Journal 22 (1991), 282–293

A watershed event for all students of mathematics is the first course in basic high school algebra. In my case, this provided an initial look at graphs, inequalities, the quadratic formula, and many other critical ideas. Somewhere near the term's end, as I remember, our teacher mentioned what sounded like the most important result of them all—the fundamental theorem of algebra. Anything with a name like that, I figured, must be (for want of a better term) *fundamental*. Unfortunately, the teacher informed us that this theorem was much too advanced to state, let alone to investigate, at our current level of mathematical development.

Fine. I was willing to wait. However, second-year algebra came and went, yet the fundamental theorem occupied only an obscure footnote from which I learned that it had something to do with factoring polynomials and solving polynomial equations. My semester in college algebra/precalculus the following year went a bit further, and I emerged vaguely aware that the fundamental theorem of algebra said that n th-degree polynomials could be factored into n (possibly complex) linear factors, and thus n th-degree polynomial equations must have n (possibly complex and possibly repeated) solutions. Of course, to that point we had done little with complex numbers and less with complex solutions of polynomial equations, so the whole business remained obscure and mysterious. Even in those pre-Watergate days, I began to sense that the mathematical establishment was engaged in some kind of cover-up to keep us ignorant of the true state of algebraic affairs.

“Oh well,” I thought, “I’m off to college, where surely I’ll get the whole story.” Four years later I was still waiting. My undergraduate mathematics training—particularly courses in linear and abstract algebra—examined such concepts as groupoids,

eigenvalues, and integral domains, but none of my algebra professors so much as mentioned the fundamental theorem. This was very unsatisfactory—a bit like reading *Moby Dick* and never encountering the whale. The cover-up had continued through college, and algebra's superstar theorem was as obscure as ever.

It was finally in a graduate school course on complex analysis that I saw a proof of this key result, and I immediately realized the trouble: the theorem really is a monster to prove in full generality, for it requires some sophisticated preliminary results about complex functions. Clearly a complete proof is beyond the reach of elementary mathematics.

So what does a faculty member do if an inquiring student seeks information about the fundamental theorem of algebra? It is hopeless to try to prove the thing for any precalculus student whose I.Q. lies on this side of Newton's; on the other hand, it would more or less continue the cover-up to avoid answering the question—to treat an inquiry about the fundamental theorem of algebra as though the student had asked something truly improper, delicate, or controversial—like a question about one's religion, or one's sex life, or even one's choice of personal computer.

Let me, then, suggest an intermediate option—something less rigorous than a grad school proof, yet something more satisfying than simply telling our inquisitive student to get lost. My suggestion is that we look back to the history of mathematics and to the work of that most remarkable of eighteenth-century mathematicians, Leonhard Euler (1707–1783). With Euler's attempted proof of the fundamental theorem of algebra from 1749, we find yet another example of the history of mathematics serving as a helpful ingredient in the successful teaching of the

subject. The reasoning is not impossibly difficult; it raises some interesting questions for further discussion; and while his is not a complete proof by any means, it does establish the result for low degree polynomials and suggests to students that this sweeping theorem is indeed reasonable.

Before addressing the subject further, we state the theorem in its modern form:

Any n th-degree polynomial with complex coefficients can be factored into n complex linear factors.

That is, if $P(z) = c_n z^n + c_{n-1} z^{n-1} + \dots + c_2 z^2 + c_1 z + c_0$, where $c_n, c_{n-1}, \dots, c_2, c_1, c_0$ are complex numbers, then there exist complex numbers $\alpha_1, \alpha_2, \dots, \alpha_n$ such that

$$P(z) = c_n(z - \alpha_1)(z - \alpha_2) \cdots (z - \alpha_n).$$

It may come as a surprise that, to mathematicians of the mid-eighteenth century, the fundamental theorem appeared in the following guise:

Any polynomial with real coefficients can be factored into the product of real linear and/or real quadratic factors.

Note that there is no mention here of complex numbers, either as the polynomial's coefficients nor as parts of its factors. For mathematicians of the day, the theorem described a phenomenon about *real* polynomials and their *real* factors.

As an example, consider the factorization

$$\begin{aligned} 3x^4 + 5x^3 + 10x^2 + 20x - 8 \\ = (3x - 1)(x + 2)(x^2 + 4). \end{aligned}$$

Here the quartic has been shattered into the product of two linear fragments and one irreducible quadratic one, and all polynomials in sight are real. The theorem stated that such a factorization was possible for any real polynomial, no matter its degree.

Anticipating a bit, we see that we can further factor the quadratic expression—provided we allow ourselves the luxury of complex numbers. That is,

$$\begin{aligned} ax^2 + bx + c &= a \left(x^2 + \frac{b}{a}x + \frac{c}{a} \right) \\ &= a \left(x - \frac{-b + \sqrt{b^2 - 4ac}}{2a} \right) \\ &\quad \times \left(x - \frac{-b - \sqrt{b^2 - 4ac}}{2a} \right) \end{aligned}$$

factors the real quadratic $ax^2 + bx + c$ into two, albeit rather unsightly, linear pieces. Of course, there is no

guarantee these linear factors are composed of *real* numbers, for if $b^2 - 4ac < 0$, we venture into the realm of imaginaries. In the specific example cited above, for instance, we get the complete factorization:

$$\begin{aligned} 3x^4 + 5x^3 + 10x^2 + 20x - 8 \\ = (3x - 1)(x + 2)(x - 2i)(x + 2i). \end{aligned}$$

This is “complete” in the sense that the real fourth-degree polynomial with which we began has been factored into the product of four *linear* complex factors, certainly as far as any factorization can hope to proceed.

It was the Frenchman Jean d’Alembert (1717–1783) who gave this theorem its first serious treatment in 1746 [5, p. 99]. Interestingly, for d’Alembert and his contemporaries the result had importance beyond the realm of algebra: its implications extended to the relatively new subject of calculus and in particular to the integration technique we now know as “partial fractions.” As an illustration, suppose we sought the indefinite integral

$$\int \frac{28x^3 - 4x^2 + 69x - 14}{3x^4 + 5x^3 + 10x^2 + 20x - 8} dx.$$

To be sure, this looks like absolute agony, as all calculus teachers will readily agree. (One would have trouble finding it in the Table of Integrals of a calculus book’s inside cover, unless the book is very thorough or its cover is very large.) This problem even gives a good workout to symbolic manipulators such as *Mathematica* (which required 50 seconds to find the antiderivative on my Mac II) and which were not available to eighteenth century mathematicians in any case.

But if, as d’Alembert claimed, the denominator could be decomposed into real linear and/or real quadratic factors, then the difficulties drop away. Here, the integrand becomes

$$\int \frac{28x^3 - 4x^2 + 69x - 14}{(3x - 1)(x + 2)(x^2 + 4)} dx.$$

We then determine its partial fraction decomposition, getting

$$\begin{aligned} \int \frac{28x^3 - 4x^2 + 69x - 14}{3x^4 + 5x^3 + 10x^2 + 20x - 8} dx \\ = \int \frac{28x^3 - 4x^2 + 69x - 14}{(3x - 1)(x + 2)(x^2 + 4)} dx \end{aligned}$$

$$\begin{aligned}
&= \int \frac{1}{3x-1} dx + \int \frac{7}{x+2} dx + \int \frac{2x-3}{x^2+4} dx \\
&= \frac{1}{3} \ln |3x-1| + 7 \ln |x+2| + \ln(x^2+4) \\
&\quad - \frac{3}{2} \tan^{-1}(x/2) + C,
\end{aligned}$$

and the antiderivative is found.

Thus, if the fundamental theorem were proved in general, we could conclude that for any $P(x)/Q(x)$ where P and Q are real polynomials, the indefinite integral $\int (P(x)/Q(x)) dx$ would exist as a combination of fairly simple functions (at least theoretically). That is, we could first perform long division to reduce this rational expression to one where the degree of the numerator was less than the degree of $Q(x)$, next we consider $Q(x)$ as the product of real linear and/or real quadratic factors; then apply the partial fraction technique to break the integral into pieces of the form

$$\int \frac{A}{(ax+b)^n} dx$$

and/or

$$\int \frac{Bx+C}{(ax^2+bx+c)^n} dx;$$

and finally determine these indefinite integrals using nothing worse than natural logarithms, inverse tangents, or trigonometric substitution. Admittedly, the fundamental theorem gives no process for finding the denominator's explicit factors; but, just as the theorem guarantees the *existence* of such a factorization, so too will the *existence* of simple antiderivatives for any rational function be established.

Unfortunately, d'Alembert's 1746 attempt to prove his theorem was unsuccessful, for the difficulties it presented were simply too great for him to overcome (see [4, pp. 196–198]). In spite of this failure, the fundamental theorem of algebra has come to be known as “d'Alembert's Theorem” (especially in France). Attaching his name to this result may seem a bit generous, given that he failed to prove it. This is a bit like designating the Battle of Waterloo as “Napoleon's Victory.”

So matters stood when Euler turned his awesome mathematical powers to the problem. At the time he picked up the scent, there was not even universal agreement that the theorem was true. In 1742, for instance, Nicholas Bernoulli had expressed to Euler his conviction that the real quartic polynomial

$$x^4 - 4x^3 + 2x^2 + 4x + 4$$

cannot be factored into the product of real linear and/or real quadratic factors in any fashion whatever [I, pp. 82–83]. If Bernoulli were correct, the game was over; the fundamental theorem of algebra would have been instantly disproved.

However, Bernoulli's skepticism was unfounded, for Euler factored the quartic into the product of the quadratics

$$x^2 - \left(2 + \sqrt{4 + 2\sqrt{7}}\right)x + \left(1 + \sqrt{4 + 2\sqrt{7} + \sqrt{7}}\right)$$

and

$$x^2 - \left(2 - \sqrt{4 + 2\sqrt{7}}\right)x + \left(1 - \sqrt{4 + 2\sqrt{7} + \sqrt{7}}\right).$$

Those with a taste for multiplying polynomials can check that these complicated factors yield the fairly innocent quartic above; far more challenging, of course, is to figure out how Euler derived this factorization in the first place. (Hint: it was not by guessing.)

By 1742, Euler claimed he had proved the fundamental theorem of algebra for real polynomials up through the sixth-degree [3, p. 598], and in a landmark 1749 article titled “Recherches sur les racines imaginaires des équations” [I, pp. 78–169], he presented his proof of the general result which we shall now examine (see also [5, pp. 100–102]). We stress again that his argument failed in its ultimate mission. That is, Euler furnished only a partial proof which, in its full generality, suffered logical shortcomings. Nonetheless, even with these shortcomings, one cannot fail to recognize the deftness of a master at work.

He began with an attack on the quartic:

Theorem. Any quartic polynomial $x^4 + Ax^3 + Bx^2 + Cx + D$ where $A, B, C,$ and D are real can be decomposed into two real factors of the second degree.

Proof. Euler first observed that the substitution $x = y - (A/4)$ reduces the original quartic into one lacking a cubic term—a so-called “depressed quartic.” Depressing an n th-degree polynomial by a clever substitution that eliminates its $(n-1)$ st-degree term is a technique whose origin can be traced to the sixteenth-century Italian mathematician Gerolamo Cardano in his successful attack on the cubic equation [3, p. 265].

With this substitution, the quartic becomes

$$\begin{aligned}
\left(y - \frac{A}{4}\right)^4 + A\left(y - \frac{A}{4}\right)^3 + B\left(y - \frac{A}{4}\right)^2 \\
+ C\left(y - \frac{A}{4}\right) + D,
\end{aligned}$$

and the only two sources of a y^3 term are

$$\left(y - \frac{A}{4}\right)^4 = y^4 - Ay^3 + \dots$$

and

$$A\left(y - \frac{A}{4}\right)^3 = A(y^3 - \dots) = Ay^3 - \dots,$$

Upon simplifying, we find that the “ y^3 ” terms cancel and there remains the promised depressed quartic in y .

Not surprisingly, there are advantages to factoring a depressed quartic rather than a full-blown one; yet it is crucial to recognize that any factorization of the depressed quartic yields a corresponding factorization of the original. For instance, suppose we were trying to factor $x^4 + 4x^3 - 9x^2 - 16x + 20$ into a product of two quadratics. The substitution $x = y - \frac{4}{4} = y - 1$ depresses this to $y^4 - 15y^2 + 10y + 24$, and a quick check confirms the factorization:

$$y^4 - 15y^2 + 10y + 24 = (y^2 - y - 2)(y^2 + y - 12).$$

Then, making the reverse substitution $y = x + 1$ yields

$$x^4 + 4x^3 - 9x^2 - 16x + 20 = (x^2 + x - 2)(x^2 + 3x - 10),$$

and the original quartic is factored as claimed.

Having reduced the problem to that of factoring depressed quartics, Euler noted that we need only consider $x^4 + Bx^2 + Cx + D$, where B, C , and D are real. At this point, two cases present themselves:

Case 1. $C = 0$.

This amounts to having a depressed quartic $x^4 + Bx^2 + D$, which is just a quadratic in x^2 . (Euler omitted discussion of this possibility, perhaps because it could be handled in two fairly easy subcases by purely algebraic means.)

First of all, suppose $B^2 - 4D \geq 0$ and apply the quadratic formula to get the decomposition into two second-degree *real* factors as follows:

$$x^4 + Bx^2 + D = \left(x^2 + \frac{B - \sqrt{B^2 - 4D}}{2}\right) \times \left(x^2 + \frac{B + \sqrt{B^2 - 4D}}{2}\right).$$

For instance, $x^4 + x^2 - 12 = (x^2 - 3)(x^2 + 4)$.

Less direct is the case where we try to factor $x^4 + Bx^2 + D$ under the condition that $B^2 - 4D < 0$.

The previous decomposition no longer works, since the factors containing $\sqrt{B^2 - 4D}$ are not real. Fortunately, a bit of algebra shows that the quartic can be written as the difference of squares and thus factored into quadratics as follows:

$$\begin{aligned} x^4 + Bx^2 + D &= (x^2 + \sqrt{D})^2 - \left(x\sqrt{2\sqrt{D} - B}\right)^2 \\ &= (x^2 + \sqrt{D} - x\sqrt{2\sqrt{D} - B}) \\ &\quad \times (x^2 + \sqrt{D} + x\sqrt{2\sqrt{D} - B}). \end{aligned}$$

A few points must be made about this factorization. First, $B^2 - 4D < 0$ implies that $4D > B^2 \geq 0$, and so the expression \sqrt{D} in the preceding factorization is indeed real. Likewise, $4D > B^2$ guarantees that $\sqrt{4D} > \sqrt{B^2}$, or simply $2\sqrt{D} > |B| \geq B$, and so the expression $\sqrt{2\sqrt{D} - B}$ is likewise real. In short, the factors above are two real quadratics, as we hoped.

For example, when factoring $x^4 + x^2 + 4$, we find $B^2 - 4D = -15 < 0$ and the formula yields $x^4 + x^2 + 4 = [x^2 - x\sqrt{3} + 2][x^2 + x\sqrt{3} + 2]$.

Case 2. $C \neq 0$.

Here Euler observed that a factorization of his depressed quartic into real quadratics—if it exists—*must* take the form

$$\begin{aligned} x^4 + Bx^2 + Cx + D \\ = (x^2 + ux + \alpha)(x^2 - ux + \beta) \end{aligned} \quad (1)$$

for some real numbers u, α , and β yet to be determined. Of course, this form is necessary since the “ ux ” in one factor must have a compensating “ $-ux$ ” in the other.

Euler multiplied out the right-hand side of (1) to get:

$$\begin{aligned} x^4 + Bx^2 + Cx + D \\ = x^4 + (\alpha + \beta - u^2)x^2 + (\beta u - \alpha u)x + \alpha\beta, \end{aligned}$$

and then equated coefficients from the first and last of these expressions to generate three equations:

$$\begin{aligned} B &= \alpha + \beta - u^2, \\ C &= \beta u - \alpha u = (\beta - \alpha)u, \text{ and} \\ D &= \alpha\beta. \end{aligned}$$

Note that B, C , and D are just the coefficients of the original polynomial, whereas u, α , and β are

unknown real numbers whose *existence* Euler had to establish.

From the first two of these we conclude that

$$\alpha + \beta = B + u^2 \quad \text{and} \quad \beta - \alpha = \frac{C}{u}.$$

It may be worth noting that since

$$0 \neq C = (\beta - \alpha)u,$$

then u itself is non-zero, so its presence in the denominator above is no cause for alarm.

If we both add and subtract these two equations, we arrive at

$$2\beta = B + u^2 + \frac{C}{u} \quad \text{and} \quad 2\alpha = B + u^2 - \frac{C}{u}. \quad (2)$$

Euler recalled that $D = \alpha\beta$ and consequently:

$$\begin{aligned} 4D &= 4\alpha\beta = (2\beta)(2\alpha) \\ &= \left(B + u^2 + \frac{C}{u}\right) \left(B + u^2 - \frac{C}{u}\right). \end{aligned}$$

In other words, $4D = u^4 + 2Bu^2 + B^2 - (C^2/u^2)$, and multiplying through by u^2 gives us

$$u^6 + 2Bu^4 + (B^2 - 4D)u^2 - C^2 = 0. \quad (3)$$

It may appear that things have gotten worse, not better, for we have traded a fourth-degree equation in x for a sixth-degree equation in u . Admittedly, (3) is also a cubic in u^2 , so we can properly conclude that there is a real solution for u^2 ; this, unfortunately, does not guarantee the existence of a *real* value for u , which was Euler's objective.

Undeterred, he noticed four critical properties of (3):

- (a) B , C , and D are known, so the only unknown here is u .
- (b) B , C , and D are real.
- (c) the polynomial is even and thus its graph is symmetric about the y -axis.
- (d) the constant term of this sixth-degree polynomial is $-C^2$.

Here Euler's mathematical agility becomes especially evident. He was considering a sixth-degree real polynomial whose graph looks something like that shown in Figure 1. This has a negative y -intercept at $(0, -C^2)$ since C is a non-zero real number. Additionally, since the polynomial is monic of even degree, its graph climbs toward $+\infty$ as u becomes unbounded in either the positive or negative direction. By a result from analysis we now

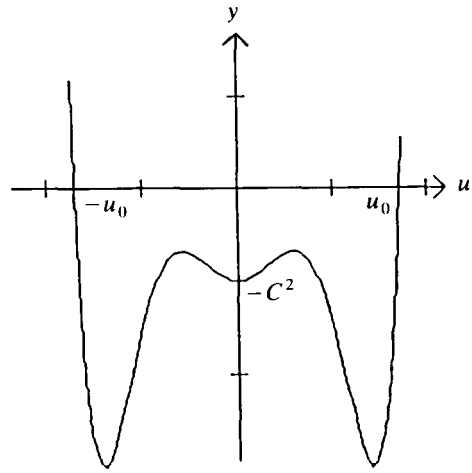


Figure 1. $y = u^6 + 2Bu^4 + (B^2 - 4D)u^2 - C^2$

call the intermediate value theorem—but which Euler took as intuitively clear—we are guaranteed the *existence* of real numbers $u_0 > 0$ and $-u_0 < 0$ satisfying this sixth-degree equation.

Using the positive solution u_0 and returning to equations in (2), Euler solved for β and α , getting real solutions

$$\beta_0 = \frac{1}{2} \left(B + u_0^2 + \frac{C}{u_0} \right)$$

and

$$\alpha_0 = \frac{1}{2} \left(B + u_0^2 - \frac{C}{u_0} \right)$$

and, since $u_0 > 0$, these fractions are well-defined.

In summary, under the case that $C \neq 0$, Euler had established the existence of real numbers u_0 , α_0 , and β_0 such that

$$x^4 + Bx^3 + Cx + D = (x^2 + u_0x + \alpha_0)(x^2 - u_0x + \beta_0).$$

We thus see that any depressed quartic with real coefficients—and by extension any real quartic at all—does have a factorization into two real quadratics, whether or not $C = 0$. Q.E.D.

At this point, Euler immediately observed, “. . . it is also evident that any equation of the fifth degree is also resolvable into three real factors of which one is linear and two are quadratic” [1, p. 95]. His reasoning was simple (see Figure 2). Any *odd-degree* polynomial—and thus any fifth-degree polynomial $P(x)$ —is guaranteed by the intermediate value theorem to have at least one real x -intercept, say at $x = a$. We then write $P(x) = (x - a)Q(x)$, where $Q(x)$ is a polynomial of the fourth degree, and the

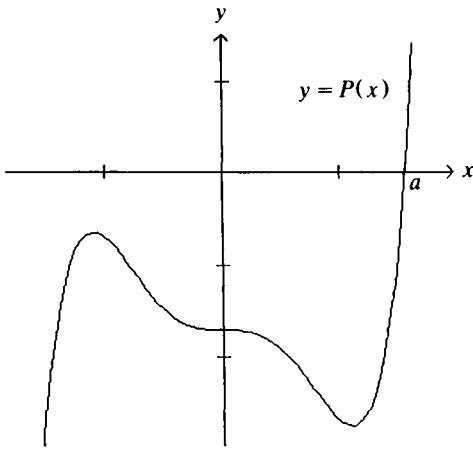


Figure 2.

previous result allows us to decompose $Q(x)$, in turn, into two real quadratic factors.

By now, a general strategy was brewing in his mind. He realized that *if* he could prove his decomposition for real polynomials of degree 4, 8, 16, 32, and in general of degree 2^n , then he could prove it for any real polynomials whatever.

Why is this? Suppose, for instance, we were trying to establish that the polynomial

$$x^{12} - 3x^9 + 5x^8 + 3x^3 - 2x + 17$$

could be factored into real linear and/or real quadratic factors. We would simply multiply it by x^4 to get

$$x^{16} - 3x^{13} + 52x^{12} + 3x^7 - 2x^5 + 17x^4.$$

Assuming that Euler had proved the 16th-degree case, he would know that this latter polynomial would have such a factorization, obviously containing the four linear factors x , x , x , and x . If we merely cancelled them out, we would of necessity be left with the real linear and/or real quadratic factors for the original 12th-degree polynomial.

And so, with typical Eulerian cleverness, he reduced the entire issue to a few simpler cases. Having disposed of the fourth-degree case, he next claimed, "Any equation of the eighth degree is always resolvable into two real factors of the fourth degree" [1, p. 99]. Since each of the fourth-degree factors was itself decomposable into a pair of real quadratics, which themselves can be broken into (possibly complex) linear factors, he would have succeeded in shattering the eighth-degree polynomial into eight linear pieces. From there he went to the 16th degree

before finally tackling the general situation, namely showing that any real polynomial of degree 2^n can be factored into two real polynomials each of degree 2^{n-1} [1, p. 105].

It was a brilliant strategy. Unfortunately, the proofs he furnished left something to be desired. As we shall see, for the higher-degree cases the arguments became hopelessly complicated, and his assertions as to the existence of *real* numbers satisfying certain equations were unconvincing. Consider, for instance, the eighth-degree case. It began in a fashion quite similar to its fourth-degree counterpart, namely by first depressing the octic and imagining that it has been factored into the two quartics:

$$\begin{aligned} x^8 + Bx^6 + Cx^5 + Dx^4 + Ex^3 + Fx^2 + Gx + H \\ = (x^4 + ux^3 + \alpha x^2 + \beta x + \gamma) \\ \times (x^4 - ux^3 + \delta x^2 + \epsilon x + \phi). \end{aligned} \quad (4)$$

One multiplies the quartics, equates the resulting coefficients with the known quantities B, C, D, \dots to get seven equations in seven unknowns, and asserts that there exist real values of $u, \alpha, \beta, \gamma, \dots$ satisfying this system.

The parallels with what he had previously done are evident. But what made this case so much less successful was Euler's admission that for equations of higher degree, "... it will be very difficult and even impossible to find the equation by which the unknown u is determined" [1, p. 97]. In short, he was unwilling or unable to solve this system explicitly for u .

Ever resourceful, Euler decided to look again at the depressed quartic in (1) for inspiration. As it turned out, an entirely different line of reasoning suggested itself, a line that he thought could be extended naturally to the eighth and higher-degree cases:

Assuming that the quartic in (1) has four roots p, q, r , and s , Euler wrote:

$$\begin{aligned} (x^2 + ux + \alpha)(x^2 - ux + \beta) \\ = x^4 + Bx^2 + Cx + D \\ = (x - p)(x - q)(x - r)(x - s), \end{aligned} \quad (5)$$

and from this factorization he drew three key conclusions.

First, upon multiplying the four linear factors on the right of (5), we see immediately that the coefficient of x^3 is $-(p+q+r+s)$; hence $p+q+r+s = 0$ since the quartic is depressed.

Second, the quadratic factor $(x^2 - ux + \beta)$ must arise as the product of two of the four linear factors. Thus, $(x^2 - ux + \beta)$ could be

$$(x - p)(x - r) = x^2 - (p + r)x + pr;$$

it could just as well be

$$(x - q)(x - r) = x^2 - (q + r)x + qr;$$

and so on. This implies that, in the first case, $u = p + r$, whereas in the second $u = q + r$. In fact, it is clear that u can take any of the $\binom{4}{2} = 6$ values

$$\begin{aligned} R_1 &= p + q, & R_2 &= r + s, & R_3 &= p + r, \\ R_4 &= q + s, & R_5 &= p + s, & R_6 &= q + r. \end{aligned}$$

Since u is an unknown having these six possible values, it must be determined by the sixth-degree polynomial

$$(u - R_1)(u - R_2)(u - R_3)(u - R_4)(u - R_5)(u - R_6).$$

This conclusion, of course, is entirely consistent with the explicit sixth-degree polynomial for u that Euler had found in (3).

But Euler made one additional observation. Because $p + q + r + s = 0$, it follows that $R_4 = -R_1$, $R_5 = -R_2$, and $R_6 = -R_3$. Hence the sixth-degree polynomial becomes

$$\begin{aligned} &(u - R_1)(u + R_1)(u - R_2) \\ &\quad \times (u + R_2)(u - R_3)(u + R_3) \\ &= (u^2 - R_1^2)(u^2 - R_2^2)(u^2 - R_3^2). \end{aligned}$$

The constant term here — which is to say, this polynomial's y -intercept — is simply

$$-R_1^2 R_2^2 R_3^2 = -(R_1 R_2 R_3)^2.$$

This constant, Euler stated, was a negative real number, again in complete agreement with his conclusions from equation (3).

To summarize, Euler had provided an entirely different argument to establish that, in the quartic case, u is determined by a $\binom{4}{2} = 6$ th-degree polynomial with a negative y -intercept. This was the critical conclusion he had already drawn, but here he drew it without *explicitly* finding the equation determining u .

The advantage of this alternate proof for the quartic case was that it could be used to analyze the depressed octic in (4). *Assuming* that the octic was decomposed into eight linear factors, Euler mimicked his reasoning above to deduce that for each different

combination of four of these eight factors, we would get a different value of u . Thus, u would be determined by a polynomial of degree $\binom{8}{4} = 70$ having a negative y -intercept. He then confidently applied the intermediate value theorem to get his desired *real* root u_0 , and from this he claimed that the other real numbers $\alpha_0, \beta_0, \gamma_0, \delta_0, \epsilon_0$, and ϕ_0 exist as well.

Euler reasoned similarly in the 16th-degree case, claiming that “. . . the equation which determines the values of the unknown u will necessarily be of the 12870th degree” [I, p. 103]. The degree of this (obviously unspecified) equation is simply $\binom{16}{8} = 12870$, as his pattern suggested. By this time, Euler's comment that it was “. . . very difficult and even impossible . . .” to specify these polynomials had become something of an understatement.

From there it was a short and entirely analogous step to the general case: that any real polynomial of degree 2^n could be factored into two real polynomials of degree 2^{n-1} . With that, his proof was finished.

Or was it? Unfortunately, his analyses of the 8th-degree, 16th-degree, and general cases were flawed and left significant questions unanswered. For instance, if we look back at the quartic in (5), how could Euler assert that it has four roots? How could he assert that the octic in (4) has eight?

More significantly, what is the nature of these supposed roots? Are they real? Are they complex? Or are they an unspecified—and perhaps entirely unimagined—new kind of number? If so, can they be added and multiplied in the usual fashion?

These are not trivial questions. In the quartic case above, for example, if we are uncertain about the nature of the roots p, q, r , and s , then we are equally uncertain about the nature of their sums R_1, R_2, R_3 . Consequently, there is no guarantee whatever that mysterious expressions such as $-(R_1 R_2 R_3)^2$ are negative real numbers. But if these y -intercepts are not negative reals, then the intermediate value arguments that Euler applied to the 8th-degree, 16th-degree, and general cases fall apart completely.

It appears, then, that Euler had started down a very promising path in his quest of the fundamental theorem. His first proof worked nicely in dealing with fourth- and fifth-degree real polynomials. But as he pursued this elusive theorem deeper into the thicket, complications involving the existence of his desired real factors became overwhelming. In a certain sense, he lost his way among the enormously high degree polynomials that beckoned him on, and his general proof vanished in the wilderness.

So even Euler suffered setbacks, a fact from which comfort may be drawn by lesser mathematicians (a category that includes virtually everybody else in history). Yet, before the dust settles and his attempted proof is consigned to the scrap heap, I think it deserves at least a modest round of applause, for it certainly bears signs of his characteristic cleverness, boldness, and mental agility as he leaps between the polynomial's analytic and algebraic properties. More to the point, the fourth- and fifth-degree arguments are understandable by good precalculus students and can give them not only a deeper look at this remarkable theorem but also a glimpse of a mathematical giant at work. For even when he stumbled, Leonhard Euler left behind signs of great insight. Such, perhaps, is the mark of genius.

Epilogue

The fundamental theorem of algebra—the result that established the complex numbers as the optimum realm for factoring polynomials or solving polynomial equations—thus remained in a very precarious state. D'Alembert had not proved it; Euler had given an unsatisfactory proof. It was obviously in need of major attention to resolve its validity once and for all.

Such a resolution awaited the last year of the eighteenth century and came at the hands of one of history's most talented and revered mathematicians. It was the 22-year old German Carl Friedrich Gauss (1777–1855) who first presented a reasonably complete proof of the fundamental theorem (see [4, p. 196] for an interesting twist on this oft-repeated statement). Gauss' argument appeared in his 1799 doctoral dissertation with the long and descriptive title, “A New Proof of the Theorem That Every Integral Rational Algebraic Function [i.e., every polynomial with real coefficients] Can Be Decomposed into Real Factors of the First or Second Degree” (see [5, pp. 115–122]). He began by reviewing past attempts at proof and giving criticisms of each. When addressing Euler's “proof,” Gauss raised the issues cited above, designating Euler's mysterious, hypothesized roots as “shadowy.” To Gauss, Euler's attempt lacked “. . . the clarity which is required in mathematics” [2, p. 491]. This clarity he attempted to provide, not only in the dissertation but in two additional proofs from 1816 and another from 1848.

As indicated by his return to this result throughout his illustrious career, Gauss viewed the fundamental theorem of algebra as a great and worthy project indeed.

We noted previously that this crucial proposition is seen today in somewhat greater generality than in the early nineteenth century, for we now transfer the theorem entirely into the realm of complex numbers in this sense: the polynomial with which we begin no longer is required to have real coefficients. In general, we consider n th-degree polynomials having complex coefficients, such as

$$z^7 + 6iz^6 - (2 + i)z^2 + 19.$$

In spite of this apparent increase in difficulty, the fundamental theorem nonetheless proves that it can be factored into the product of (in this case seven) linear terms having, of course, complex coefficients. Interestingly, modern proofs of this result almost never appear in algebra courses. Rather, today's proofs rest upon a study of the *calculus* of complex numbers and thus move quickly into the realm of genuinely advanced mathematics (just as my high school algebra teacher had so truthfully said).

And so, we reach the end of our story, a story that can be a valuable tale for us and our students. It addresses an oft-neglected theorem of much importance; it allows the likes of Jean d'Alembert, Leonhard Euler, and Carl Friedrich Gauss to cross the stage; and it gives an intimate sense of the historical development of great mathematics in the hands of great mathematicians.

References

1. Leonhard Euler, Recherches sur les racines imaginaires des équations, *Mémoires de l'académie des sciences de Berlin* (5) (1749), 1751, 222–288 (*Opera Omnia* (1) 6, 78–147).
2. J. Fauvel and J. Gray, *The History of Mathematics: A Reader*, Macmillan, London, 1987.
3. Morris Kline, *Mathematical Thought from Ancient to Modern Times*, Oxford University Press, New York, 1972.
4. John Stillwell, *Mathematics and its History*, Springer-Verlag, New York, 1989.
5. Dirk Struik (Ed.), *A Source Book in Mathematics: 1200–1800*, Princeton University Press, 1986.

Euler and Differentials

ANTHONY P. FERZOLA

College Mathematics Journal 25 (1994), 102–111

Two recent articles by Dunham [5] and Flusser [10] have presented examples of Leonhard Euler's work in algebra. Both papers are a joy to read; watching Euler manipulate and calculate with incredible facility is a pleasure. A modern mathematician can see the logical flaws in some of the arguments, yet at the same time be aware that the mind behind it all is that of a unique master.

These two articles reminded me how much fun it is to read Euler. In researching the evolution of the differential a few years ago, I found the work of Euler refreshingly different from that of other seventeenth- and eighteenth-century mathematicians. One can read about Euler's use and misuse of infinite series in most histories of mathematics (e.g. [2, pp. 486–490]). This paper offers a glimpse at how Euler used infinitesimals and infinite series to compute differentials for the elementary functions encountered in a typical undergraduate calculus sequence. I hope the reader of this brief survey of Euler's work with differentials will seek out original sources such as [8] and [9]. As Harold Edwards [7] has cogently argued, we have much to learn from reading the masters.

Euler and the 18th century

Euler (1707–1783) was the most prolific and one of the most influential mathematicians who ever lived. He made major contributions to both pure and applied mathematics and his collected works amount to over 70 volumes. So strong was his influence that historians like Boyer [2] and Edwards [6] refer to the eighteenth century as the Age of Euler.

Euler made the function concept fundamental in analysis. He saw a function as both any quantity depending on variables and also as any algebraic

combination of constants and variables (including infinite sums or products). This is obviously not a modern definition of a function. Still, Euler used his function concept to maximal advantage. As we examine some of Euler's computations, keep in mind the immense insight and unity he achieved with the function approach — a point of view we now take for granted.

In his *Introductio in analysin infinitorum* (1748), one sees the first systematic interpretation of logarithms as exponents. Prior to Euler, logarithms were typically viewed as terms of an arithmetic series in one-to-one correspondence with terms of a geometric series [3]. Euler viewed trigonometric functions as numerical ratios rather than as ratios of line segments. He also studied properties of the elementary transcendental functions by the frequent use of their infinite series expansions [6, p. 270]. Euler often used infinite series indiscriminately, without regard to questions of convergence.

Euler's understanding and use of differentials within the framework of functions is the focus of this paper. Before presenting his work, a word about the differential before Euler.

For Leibniz (1646–1716) the differentials dx and dy were, as the name suggests, (infinitesimal) differences in the abscissa x and the ordinate y , respectively [4, pp. 70–76]. The infinitesimal was considered to be a number smaller than any positive number. The omission of the “even smaller” higher-order infinitesimals such as $(dx)^2$ or $dx dy$, which were deemed negligible relative to dx and dy , was basic to his methods. So powerful were the notation and methods that the differential calculus was truly a *differential* calculus for nearly one and a half centuries: The differential (and not the derivative) was the main object of study.

Leibniz gave other interpretations of the differential, but the mathematicians working in the early eighteenth century tended to favor Leibniz's formulation of a differential as an infinitesimal. It appears in the work of Johann Bernoulli (1667–1748) and in the first calculus textbook, *Analyse des infiniment petits pour l'intelligence des lignes courbes* (1696), which was written by L'Hôpital and which made free use of Bernoulli's ideas (see [18, p. 315]). Euler was one of Bernoulli's pupils.

Many of Euler's results and infinite series discussed below were known to Newton, Leibniz, Bernoulli, and others. Euler's work with differentials is unique, however, in his definition of infinitesimals as absolute zeros and in his heavy reliance on infinite series to develop his differential calculus.

Differentials as absolute zeros

In his *Institutiones calculi differentialis* (1755), Euler stated: "To those who ask what the infinitely small quantity in mathematics is, we answer it is actually equal to zero" [18, p. 384]. Euler felt that the view of the infinitesimal as zero adequately removed the mystery and ambiguity of statements such as "The infinitesimal is smaller than any given quantity" or the postulate of Johann Bernoulli that "Adding an infinitesimal to a quantity leaves the quantity unchanged." Euler then said that the quotient $0/0$ can actually take on any value because

$$n \cdot 0 = 0$$

for all real n and therefore, he concluded,

$$\frac{n}{1} = \frac{0}{0}, \quad (1)$$

He noted that if two zeros can have an arbitrary ratio, then different symbols should be used for the zero in the numerator and the zero in the denominator of the fraction on the right-hand side of equation (1). It is here that Euler introduced the Leibnizian notation of differentials.

Euler denoted an infinitely small quantity by dx . Here $dx = 0$ and $adx = 0$ for any finite quantity a . But for Euler these two zeros are different zeros that cannot be confused when the ratio $adx/dx = a$ is investigated [18, p. 385]. In a similar way dy/dx can denote a finite ratio even though dx and dy are zero. "Thus for Euler the calculus was simply the determination of the ratios of evanescent increments—a heuristic procedure for finding the value of the expression $0/0$ " [1].

The neglect of higher-order infinitesimals was also explained employing quotients. Noting that $dx = 0$ and $(dx)^2 = 0$, where $(dx)^2$ is a zero (or infinitesimal) of second order, Euler reasoned that

$$dx + (dx)^2 = dx$$

because

$$\frac{dx + (dx)^2}{dx} = 1 + dx = 1.$$

By the same reasoning, Euler established that

$$dx + (dx)^{n+1} = dx$$

for all $n > 0$. The omission of higher-order differentials was frequently utilized by Euler in finding the differential dy , where y is a function of x .

Computations with elementary functions

The computations discussed in this section are all found in Euler's *Institutiones calculi differentialis*. Their most noteworthy feature is the use of power series expressions for functions from the outset, with no mention of questions of convergence. Thus, whereas in modern textbooks the justification of such infinite series expansions is an advanced topic in differential calculus, for Euler they were the foundation for the calculation of derivatives.

To find dy if $y = x^n$ (n any real number), Euler used the binomial expansion [9, p. 99]. If x is increased by an infinitesimal amount dx , then y experiences a change of dy where

$$\begin{aligned} dy &= (x + dx)^n - x^n \\ &= nx^{n-1}dx + \frac{n(n-1)}{1}x^{n-2}(dx)^2 + \dots \\ &= nx^{n-1}dx \end{aligned}$$

upon the omission of the higher-order infinitesimals $(dx)^2$, etc. Newton and Leibniz did similar computations for finding the derivative of $y = x^n$, Leibniz using a comparable differential argument while Newton worked with fluxions [6, p. 192]. Within the rigorous context of

$$\lim_{\Delta x \rightarrow 0} \frac{(x + \Delta x)^n - x^n}{\Delta x}$$

we all use the essence of this computation (for positive integer powers of x) in our first semester calculus courses.

Euler derived the product rule as follows:

$$\begin{aligned} d(pq) &= (p + dp)(q + dq) - pq \\ &= pdq + qdp + dpdq \\ &= pdq + qdp \end{aligned}$$

where the last step is due to the omission of the higher-order infinitesimal $dpdq$.

Similar computations were done by Leibniz [4, p. 143]. This argument is analogous to the proof of the product rule still found in a few present-day textbooks (e.g., [12]).

Euler's derivation of the quotient rule is unique in its use of a geometric series [9, p. 103]:

$$\begin{aligned} \frac{1}{q + dq} &= \frac{1}{q} \left(\frac{1}{1 + dq/q} \right) \\ &= \frac{1}{q} \left(1 - \frac{dq}{q} + \frac{dq^2}{q^2} - \dots \right) \\ &= \frac{1}{q} - \frac{dq}{q^2}. \end{aligned}$$

Then

$$\begin{aligned} d\left(\frac{p}{q}\right) &= \frac{p + dp}{q + dq} - \frac{p}{q} \\ &= (p + dp) \frac{1}{q + dq} - \frac{p}{q} \\ &= (p + dp) \left(\frac{1}{q} - \frac{dq}{q^2} \right) - \frac{p}{q} \\ &= \frac{dp}{q} - \frac{pdq}{q^2} \\ &= \frac{qdp - pdq}{q^2}. \end{aligned}$$

In chapter 6, Euler found the differentials of transcendental functions. For computing the differential of the natural logarithm (which he denoted by the single letter “ ℓ ” but which we will denote by the usual “ \log ”), Euler used Mercator's series [9, p. 122]:

$$\log(1 + z) = z - \frac{z^2}{2} + \frac{z^3}{3} - \dots$$

Given $y = \log(x)$ then

$$\begin{aligned} dy &= \log(x + dx) - \log(x) \\ &= \log\left(1 + \frac{dx}{x}\right) \\ &= \frac{dx}{x} - \frac{(dx)^2}{2x^2} + \frac{(dx)^3}{3x^3} - \dots \\ &= \frac{dx}{x}. \end{aligned}$$

To illustrate the chain rule, Euler did many examples. For instance, if $y = \log(x^n)$ then letting $p = x^n$ yields $y = \log(p)$, which implies that $dy = dp/p$ where $dp = nx^{n-1}dx$. Thus $dy = ndx/x$.

Euler's computation of dy for $y = \log(x)$ can be found in a modern nonstandard analysis text [15, p. 65]. This may seem unremarkable since nonstandard analysis was developed by Abraham Robinson in the mid-twentieth century to place the notion of infinitesimals and their manipulation on solid logical ground. In fact, it is rare to find nonstandard analysis arguments that are exactly like Euler's, because nonstandard analysis arguments are rarely done in the context of infinite series (see [11] and [16]).

As an example of Euler's work with trigonometric functions, consider the computation of dy for $y = \sin x$ [9, p. 132]. For this purpose he explicitly used the sine and cosine series

$$\sin x = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \dots \tag{2}$$

$$\cos x = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \dots \tag{3}$$

to show that $\sin(dx) = dx$ and $\cos(dx) = 1$. He obtained these results by substituting dx into (2) and (3) and ignoring higher-order differentials. He also employed the trigonometric identity

$$\sin(a + b) = \sin a \cos b + \sin b \cos a. \tag{4}$$

Thus, using (4):

$$\begin{aligned} dy &= \sin(x + dx) - \sin x \\ &= \sin x \cos dx + \sin dx \cos x - \sin x \\ &= \sin x + \cos x dx - \sin x \\ &= \cos x dx. \end{aligned}$$

This is the most beautifully efficient computation of all those presented, especially when compared to the usual limit computation of the derivative of $y = \sin x$. There one needs to work as follows:

$$\begin{aligned} &\lim_{\Delta x \rightarrow 0} \frac{\sin(x + \Delta x)}{\Delta x} \\ &= \lim_{\Delta x \rightarrow 0} \frac{\sin x \cos(\Delta x) + \sin(\Delta x) \cos x - \sin x}{\Delta x} \\ &= \cos x \lim_{\Delta x \rightarrow 0} \frac{\sin(\Delta x)}{\Delta x} + \sin x \lim_{\Delta x \rightarrow 0} \frac{\cos(\Delta x) - 1}{\Delta x}. \end{aligned}$$

Then $y' = \cos x$ is obtained using two limits (which must be proven):

$$\lim_{\Delta x \rightarrow 0} \frac{\sin x}{x} = 1 \tag{5}$$

and

$$\lim_{\Delta x \rightarrow 0} \frac{\cos x - 1}{x} = 0.$$

The first of these limits is captured in Euler's equation $\sin(dx) = dx$. The second limit is comparable to Euler's equation $\cos(dx) = 1$ or $\cos(dx) - 1 = 0$. Although Euler's derivation is computationally more compact than the standard modern approach, the latter is logically sound. Any method for differentiating the sine function must deal in particular with (5). This is proven geometrically, since in the standard modern approach one defines at the outset the geometric meaning of the trigonometric functions (i.e., cosine and sine parametrize the unit circle). The proof of (5) is relatively easy when compared to the difficulty involved in showing the geometric meaning of the functions Euler defined (without regard to questions of convergence) as the sums of the power series (2) and (3).

In Euler's three-volume *Institutiones calculi integralis* (1768–1770), he defined integration, like Leibniz and Johann Bernoulli, as the formal inverse of the differential. He used the integral symbol and wrote, for example,

$$\begin{aligned} \int nx^{n-1} dx &= x^n, \\ \int dx/x &= \log x, \\ \int \cos x dx &= \sin x, \end{aligned}$$

all plus or minus an appropriate constant. The first volume of this work reads like a modern calculus textbook chapter on techniques of integration. Integration by substitution, by parts, by partial fractions, and by trigonometric substitution are all illustrated in a logical and systematic way. Undoubtedly, Euler's well-organized and all-encompassing use of differentials in a function context did much to solidify the popularity of the differential and integral notations on the Continent.

The total differential

Euler's *Institutiones calculi differentialis* was the first systematic exposition of the calculus of functions of several variables. He understood a function of n variables to be any finite or infinite expression involving these variables. As soon as he introduced these functions, Euler addressed the question of the relationship among the differentials of all the variables involved.

He obtained the result that if

$$V = f(x, y, z)$$

then

$$dV = pdx + qdy + rdz,$$

where p , q , and r are all functions of x , y , and z [9, pp. 144–145]. He arrived at this formula in an interesting way. If X is a function of x alone and is increased by an infinitesimal amount dx , then

$$dX = Pdx$$

by the usual one-variable argument. Similarly, if Y and Z are functions of y alone and z alone respectively, then

$$dY = Qdy \quad \text{and} \quad dZ = Rdz.$$

If $V = X + Y + Z$ (i.e., a special function of three variables), then

$$dV = dX + dY + dZ = Pdx + Qdy + Rdz.$$

If $V = XYZ$, then

$$dV = (X + Pdx)(Y + Qdy)(Z + Rdz) - XYZ.$$

This simplifies (upon omission of higher-order differential terms such as $ZPQdx dy$) to

$$dV = YZPdx + XZQdy + XYRdz.$$

From these two examples, Euler expected that any algebraic expression of x , y , and z has differential

$$dV = pdx + qdy + rdz \quad (6)$$

because a function of three variables can be thought of as a sum of products of these variables. He generalized the result for any number of variables [9, p. 146]. Later in the same work, he addressed the concept of partial differentiation [9, pp. 156–157]. If y and z are held constant, then by equation (6)

$$dV = pdx$$

as there is no change in y or z . (Notice how, for Euler, no change in y is not the same as saying dy is the infinitesimal change in y , even though he defined infinitesimals as being zero.) He then wrote

$$p = (dV/dx),$$

where the parentheses about the quotient remind one that p equals the differential of V (with only the

x being variable) divided by dx . Similar meanings apply to $q = (dV/dy)$ and $r = (dV/dz)$. This was Euler’s notation and understanding of the concept of partial derivatives. The current symbol ∂ dates from the 1840’s [14]. Obviously, (6) becomes

$$dV = (dV/dx)dx + (dV/dy)dy + (dV/dz)dz,$$

although Euler did not explicitly write this.

It is worth noting that Euler’s exposition of differentials for functions of several variables immediately followed his work with differentials for functions of one variable. Exploring the differential calculus for both single and multivariable functions before passing on to integration is an old idea which I think has merit. It gives the calculus sequence a stronger focus and unity, by concentrating effort on one basic concept (the derivative) in various settings before moving on to its inverse. A recent textbook by Small and Hosack [17] takes this approach. Perhaps we will see more of this, especially since computer algebra systems such as *Derive*, *Maple*, and *Mathematica* have taken the pain out of such tasks as surface sketching.

Differentials in multiple integrals

Euler frequently let his readers in on his thought processes, even when the procedures seemed fruitless. This was mathematics being done for all to see, not a slick modern textbook treatment. There was no taking down the scaffolding à la Gauss.

Euler, in *De formulis integralibus duplicatis* (1769), gave one of the first clear discussions of double integrals. In the first half of the eighteenth century, $\iint f(x, y) dx dy$ denoted the solution of $\partial^2 z / \partial x \partial y = f(x, y)$ obtained by antidifferentiation. Euler supplemented this by providing a (thoroughly modern) procedure for evaluating definite double integrals over a bounded domain R enclosed by arcs in the xy plane. Euler used iterated integrals:

$$\iint_R f(x, y) dx dy = \int_a^b dx \int_{f_1(x)}^{f_2(x)} z dy,$$

where $z = f(x, y)$. For $z > 0$, Euler saw this as a volume, since $\int z dy$ gives the area of a “slice” (parallel to the y -axis) of the three-dimensional region above R and under $z = f(x, y)$, and the following integration with respect to x “adds up the slices” to yield the volume [8, p. 293]. This is perhaps the first time Leibniz’s powerful differential notation was used in tandem with a volume argument

employing Cavalieri’s method of indivisibles [2, p. 361].

Euler also interpreted $dx dy$ as an “area element” of R . That is, R is made up of an infinite set of infinitesimal area elements $dx dy$. This is most clearly seen when Euler attempted to change variables [8, pp. 302–303]. And it was here that Euler ran into difficulties.

He reasoned that if $dx dy$ is an area element and we change variables via the transformation

$$\begin{aligned} x &= x(t, v) = a + mt + v\sqrt{1 - m^2} \\ y &= y(t, v) = b + t\sqrt{1 - m^2} - mv \end{aligned}$$

(a translation by the vector (a, b) , a clockwise rotation through the angle α , where $\cos \alpha = m$, and a reflection through the x -axis), then $dx dy$ should equal $dt dv$. But

$$\begin{aligned} dx &= m dt + dv\sqrt{1 - m^2}, \\ dy &= dt\sqrt{1 - m^2} - m dv, \end{aligned}$$

and multiplication gives

$$\begin{aligned} dx dy &= m\sqrt{1 - m^2} (dt)^2 \\ &+ (1 - 2m^2) dt dv - m\sqrt{1 - m^2} (dv)^2. \end{aligned}$$

Euler rejected this as wrong and meaningless. (How many calculus students wonder, explicitly or implicitly, why we cannot just multiply the differential forms for dx and dy ?) Euler decided to attack the problem in a formal non-geometric way, not using area elements but rather by changing variables one at a time (for details, see [13]). In this way he arrived at the correct general result:

$$\begin{aligned} \iint f(x, y) dx dy &= \iint f(x(t, v), y(t, v)) \left| \frac{\partial(x, y)}{\partial(t, v)} \right| dt dv. \end{aligned}$$

In 1899, another great mathematician with a computational flair, Élie Cartan, arrived at the straightforward multiplicative result Euler sought, by using Grassmann’s exterior product with differential forms. This is a formal product where the usual distributive laws hold but with the conditions that

$$dx dx = dy dy = 0 \quad \text{and} \quad dx dy = -dy dx$$

(see [17, p. 514], and [13]). Thus, for Euler’s dif-

ferentials

$$\begin{aligned}
 dx dy &= \left(m dt + dv \sqrt{1 - m^2} \right) \\
 &\quad \times \left(dt \sqrt{1 - m^2} - m dv \right) \\
 &= dt dt m \sqrt{1 - m^2} - m^2 dt dv \\
 &\quad + (1 - m^2) dv dt - dv dv m \sqrt{1 - m^2} \\
 &= -m^2 dt dv + dv dt (1 - m^2) \\
 &= -m^2 dt dv - dt dv (1 - m^2) \\
 &= -dt dv.
 \end{aligned}$$

The minus sign appears because the transformation (involving a reflection) does not preserve orientation. In general, given any transformation from the tv -plane to the xy -plane, the exterior product yields

$$dx dy = \frac{\partial(x, y)}{\partial(t, v)} dt dv.$$

Conclusion

Even in this rudimentary survey of Euler's work with differentials in calculus, it is fascinating to watch a genius grapple with an ambiguous concept (infinitesimal) and attempt to clarify it (absolute zero) — however flawed the attempt. Reading Euler has enriched my teaching of the calculus by keeping me mindful that my students are tackling a subject whose foundations humbled the greatest minds of the past. Even the seemingly fruitless paths can be instructive, as we have seen. It took mathematicians about 150 years to come up with the exterior product for differential forms that Euler needed for the change of variables formula in multiple integrals. How many other Eulerian dead ends may be worth pursuing? Again, the advice of Harold Edwards [7] points the way for the teacher and the researcher: "Read the masters!"

References

1. C. Boyer, *The History of Calculus*, Dover, New York, 1959, p. 244.
2. C. Boyer, *A History of Mathematics*, Wiley, New York, 1968,
3. F. Cajori, *A History of Mathematics*, Macmillan, London, 1931, p. 235.
4. J. Child, *The Early Mathematical Manuscripts of Leibniz*, Open Court, Chicago, 1920.
5. W. Dunham, Euler and the fundamental theorem of algebra, *College Mathematics Journal* 22 (1991) 282–293.
6. C. H. Edwards, Jr., *The Historical Development of the Calculus*, Springer-Verlag, New York, 1979.
7. H. M. Edwards, Read the masters!, in L. A. Steen, ed., *Mathematics Tomorrow*, Springer-Verlag, New York, 1981.
8. L. Euler, *De formulis integralibus duplicatis*, in A. Gutzmer, ed., *Opera Omnia: L. Euler*, Vol. 17, Series 1, B. G. Teubner, Leipzig, 1915.
9. —, *Institutiones calculi differentialis*, in C. Kowalewski, ed., *Opera Omnia: L. Euler*, Vol. X, Series 1, B. G. Teubner, Leipzig, 1913.
10. P. Flusser, Euler's amazing way to solve equations, *Mathematics Teacher* 85 (1992) 224–227.
11. J. M. Henle and E. M. Kleinberg, *Infinitesimal Calculus*, MIT Press, Cambridge, 1979.
12. D. Hughes-Hallett, A. M. Gleason, et al., *Calculus: Preliminary Edition*, Wiley, New York, 1992, p. 239.
13. V. Katz, Change of variables in multiple integrals: Euler to Cartan, *Mathematics Magazine* 55 (1982) 3–11.
14. —, The history of differential forms from Clairaut to Poincare, *Historia Mathematica* 8 (1981) 161–188, p. 161.
15. A. Robert, *Nonstandard Analysis*, Wiley, New York, 1988, p. 65.
16. A. Robinson, *Non-Standard Analysis*, North-Holland, Amsterdam, 1970.
17. D. B. Small and J. M. Hosack, *Calculus: An Integrated Approach*, McGraw-Hill, New York, 1990.
18. D. J. Struik, *A Source Book in Mathematics 1200–1800*, Harvard University Press, Cambridge, 1969.

Euler and Quadratic Reciprocity

HAROLD M. EDWARDS

Mathematics Magazine 56 (1983), 285–291

In a letter to Goldbach bearing the date 28 August 1742, Euler described a property of positive whole numbers that was to play a central role in the history of the theory of numbers. (The original is a mixture of Latin and German, which I have translated into English as best I can. The letter can be found in [2] or [3].)

Whether there are series of numbers which either have no divisors of the form $4n + 1$, or which even are prime, I very much doubt. If such series could be found, however, one could use them to great advantage in finding prime numbers.

By the way, the prime divisors of all series of numbers which are given by the formula $\alpha xx \pm \beta yy$ show a very orderly pattern which, although I have no demonstration of it as yet, seems to be completely correct. For this reason I take the liberty of communicating to Your Excellency a few such theorems; from these, infinitely many others can be derived.

I. If x and y are relatively prime, the formula $xx + yy$ has no prime divisors other than those contained in the form $4n + 1$, and these prime numbers are themselves all contained in the form $xx + yy$. I put this known theorem at the beginning in order to make the connection of the others more apparent.

II. The formula $2xx + yy$ has no prime divisors other than those contained in the form $8n + 1$ or $8n + 3$. And whenever $8n + 1$ or $8n + 3$ is prime, it is the sum of a square and twice a square, that is, it is of the form $2xx + yy$.

III. The formula $3xx + yy$ has no prime divisors other than those contained in the forms $12n + 1$ and $12n + 7$ (or the single form $6n + 1$).

And whenever $6n + 1$ is a prime number it is contained in the form $3xx + yy$.

IV. The formula $5xx + yy$ has no prime divisors other than those contained in the forms $20n + 1$, $20n + 3$, $20n + 7$, $20n + 9$, and every prime number contained in one of these four forms is itself a number of the form $5xx + yy$.

V. The formula $6xx + yy$ has no prime divisors other than those contained in one of the four forms $24n + 1$, $24n + 5$, $24n + 7$, $24n + 11$, and every prime number contained in one of these forms is itself a number of the form $6xx + yy$.

VI. The formula $7xx + yy$ has no prime divisors other than those contained in one of the 6 forms $28n + 1$, $28n + 9$, $28n + 11$, $28n + 15$, $28n + 23$, $28n + 25$ (or in one of the three $14n + 1$, $14n + 9$, $14n + 11$), and every prime number contained in one of these forms is itself a number of the form $7xx + yy$.

...

From this it is thus clear that the expression $pxx + yy$ can have no prime divisors other than those contained in a certain number of forms of the type $4pn + s$, where s represents some numbers which, although they appear to have no particular order, actually proceed according to a very beautiful rule, which is clarified by these theorems:

VII. If a prime number of the form $4pn + s$ is a divisor of the formula $pxx + yy$ then likewise every prime number contained in the general form $4pn + s^k$ will be a divisor of the formula $pxx + yy$ and indeed will itself be a number of the form $pxx + yy$. For example, because a prime number $28n + 9$ is a number of the form $7xx + yy$ [$37 = \text{prime} = 28 \cdot 1 + 9 =$

$7 \cdot 4 + 9]$ prime numbers $28n + 81$ ($28n + 25$) and $28n + 729$ ($28n + 1$) are indeed numbers of the form $7xx + yy$ [53 and 29].

VIII. If two prime numbers $4pn + s$ and $4pn + t$ are divisors of the formula $pxx + yy$ then every prime number of the form $4pn + s^k t^j$ is also a number of the form $pxx + yy$.

Thus when one has found a few prime divisors of such an expression $pxx + yy$ one can easily find all possible divisors using these theorems. For example, let $13xx + yy$ be the given formula, which includes the numbers 14, 17, 22, 29, 38, 49, 62, etc. Thus 1, 7, 11, 17, 19, 29, 31 are prime numbers which divide the formula $13xx + yy$. Therefore all prime numbers of the forms $52n + 1$, $52n + 7$, $52n + 11$ etc. can be divisors of $13xx + yy$. But the formula $52n + 7$ gives, by Theorem VII, also these $52n + 49$, $52n + 343$ (or $52n + 31$), $52n + 7 \cdot 31$, or $52n + 9$, further $52n + 7 \cdot 9$, or $52n + 11$, further $52n + 7 \cdot 11$, or $52n + 25$, further $52n + 7 \cdot 25$, or $52n + 19$, further $52n + 7 \cdot 19$, or $52n + 29$, further $52n + 7 \cdot 29$, or $52n + 47$, further $52n + 7 \cdot 47$, or $52n + 17$, further $52n + 7 \cdot 17$, or $52n + 15$, further $52n + 7 \cdot 15$, or $52n + 1$ and at this point the numbers cease to be different which when added to $52n$ give prime numbers of the form $13xx + yy$. Thus from the single fact that 7 can be a divisor of the form $13xx + yy$ the last two theorems imply that all prime numbers of any of the forms

$$\begin{aligned} &52n + 1; \quad 52n + 31; \quad 52n + 25; \quad 52n + 47 \\ &52n + 7; \quad 52n + 9; \quad 52n + 19; \quad 52n + 17 \\ &52n + 49; \quad 52n + 11; \quad 52n + 29; \quad 52n + 15 \end{aligned}$$

have the form $13xx + yy$ and also can be divisors of such numbers $13xx + yy$, and also more formulas can not be derived using the theorems. From this it is known that no prime number can be a divisor of the form $13xx + yy$ other than those contained in the 12 formulas that have been found. Now every prime number of the form $4pn + 1$ can be a divisor of $pxx + yy$. From this, beautiful properties can be derived, as, for example, because 17 is prime and also of the form $2xx + yy$ it follows that whenever $17^m \pm 8n$ is prime it must also be of the form $2xx + yy$. And when $17^m \pm 8n$ is a number of the form $2xx + yy$ which admits no divisors of this form, it is certainly a prime number.

The same situation occurs with the divisors of the forms $pxx - yy$ or $xx - pyy$, which, when

they are prime, must be contained in the form $4np \pm s$, where s represents certain determined numbers. Namely, in a few cases, one will have

1. All prime divisors of the form $xx - yy$ contained in the form $4n \pm 1$, which is clear.

2. All prime divisors of the form $2xx - yy$ contained in the form $8n \pm 1$.

Coroll. Therefore a prime number of the form $8n \pm 3$ is not a number of the form $2xx - yy$.

3. All prime divisors of the form $3xx - yy$ contained in the form $12n \pm 1$.

4. All prime divisors of the form $5xx - yy$ contained in either the form $20n \pm 1$ or the form $20n \pm 9$ (or in the single one $10n \pm 1$).

etc.

And if a prime number $4pn + s$ divides the form $pxx - yy$ or $xx - pyy$, then $\pm 4np \pm s^k$ will itself be of the form $pxx - yy$ or $xx - pyy$, whenever it is prime. If two prime numbers s and t are numbers of the form $pxx - yy$, then whenever $4np \pm s^\mu t^\nu$ is prime it will also be a number of the form $pxx - yy$. Thus, because 7 and 17 are prime numbers and of the form $2xx - yy$, $\pm 8n \pm 7^\mu 17^\nu$ will also be of this form whenever it is prime. Let $\mu = 1$, $\nu = 1$, so $7 \cdot 17 = 119$ and $119 + 8 = 127 = \text{prime}$, and consequently $127 = 2xx - yy = 2 \cdot 64 - 1$. From this it is now clear that it is not possible to find sequences of numbers of the type $pxx \pm qyy$ which do not admit divisors of the form $4n + 1$.

But I am convinced that I have not exhausted this material, rather, that there are countless wonderful properties of numbers to be discovered here, by means of which the theory of divisors could be brought to much greater perfection; and I am convinced that if Your Excellency were to consider this subject worthy of some attention He would make very important discoveries in it. The greatest advantage would show itself, however, when one could find proofs for these theorems.

This passage is vintage Euler in that the basic idea is an insight so profound that it is crucial to much of algebraic number theory, yet at the same time many of the individual statements are patently false. The last statement of Theorem IV, for example, is clearly wrong. Not only is it not true that *all* prime numbers of the form $20n + 3$ are of the form $5x^2 + y^2$, but *no* prime numbers $20n + 3$ are $5x^2 + y^2$. To prove this

it suffices to note that, since p is to be odd, x and y must have opposite parity, that is, either $x = 2j + 1$, $y = 2k$ or $x = 2c$, $y = 2d + 1$. In the first case

$$\begin{aligned} 5x^2 + y^2 &= (4 + 1)(4j^2 + 4j + 1) + 4k^2 \\ &= 4(4j^2 + 4j + 1 + j^2 + j + k^2) + 1 \end{aligned}$$

and in the second case

$$5x^2 + y^2 = 4(5c^2 + d^2 + d) + 1,$$

so in either case p is 1 more than a multiple of 4 and cannot have the form $4n + 3$, much less the form $20n + 3$, or the form $20n + 4 + 3$.

Fortunately, the letter to Goldbach is only the first of many passages in his known writings where Euler deals with this subject, and in later versions the obvious mistakes are corrected. For example, in his main exposition [2] of these ideas he corrects the second part of Theorem IV to say that if p is a prime of the form $p = 20n + 1$ or $20n + 9$ then $p = 5x^2 + y^2$, and if it is a prime of the form $20n + 3$ or $20n + 7$ then $2p = 5x^2 + y^2$. (Examples: $2 \cdot 3 = 5 \cdot 1^2 + 1^2$, $2 \cdot 7 = 5 \cdot 1^2 + 3^2$, $2 \cdot 23 = 5 \cdot 3^2 + 1^2$, $2 \cdot 43 = 5 \cdot 1^2 + 9^2$, $2 \cdot 47 = 5 \cdot 3^2 + 7^2$.) As restated, the theorem is correct and definitely not easy to prove.

The style of the corrected exposition [1] is similar to the letter above in that Euler first states a number of special theorems—covering the prime divisors of $a^2 + Nb^2$ (a, b relatively prime) for $N = 1, 2, 3, 5, 7, 11, 13, 17, 19, 6, 10, 14, 15, 21, 35, 30$ —before he states general theorems. This style has the advantage that the reader, far from having to struggle with the meaning of the general theorem, has probably become impatient with the special cases and has already made considerable progress toward guessing what the general theorem will be. Such a style is not appropriate to the sort of short note I am writing, however, and I will skip to the general case. Moreover, I will state it much more succinctly than Euler does.

Theorem. *Let N be a given positive integer. Then there is a list s_1, s_2, \dots, s_m of positive integers less than $4N$ and relatively prime to $4N$ with the following properties:*

(1) *Any odd prime number p which divides a number of the form $a^2 + Nb^2$ without dividing either a or Nb is of the form $p = 4Nn + s_i$, for some s_i in the list.*

(2) *Every prime number of the form $p = 4Nn + s_i$ for some s_i in the list divides a number of the form $a^2 + Nb^2$ without dividing either a or Nb .*

(3) *If s_i and s_j are in the list and if $s_i s_j = 4Nn + s$, $0 < s < 4N$, then s is in the list.*

(4) *If x is any integer less than $4N$ and relatively prime to $4N$ then either x or $4N - x$, but not both, are in the list.*

For example, when $N = 13$, the list contains the 12 numbers 1, 7, 49, 31, 9, 11, 25, 19, 29, 47, 17, 15, that Euler gave in his letter to Goldbach. Property (4) becomes clearer if one writes $-x$ in place of $4N - x$ when $2N < 4N - x < 4N$ and reorders the list in order of the size of the absolute values of the entries. In the case $N = 13$ this gives 1, -3 , -5 , 7, 9, 11, 15, 17, 19, -21 , -23 , 25, and in the general case it gives (by (4)) a list of the positive integers x less than $2N$ and relatively prime to $2N$ with a sign assigned to each. To see that property (3) holds in the case $N = 13$ it suffices to note that Euler, in the letter, derived his list 1, 7, 49, 31, ... by repeatedly multiplying by 7 and removing multiples of 52. Thus, in the case $N = 13$, the numbers s_i in the list are determined by $7^i = 52n_i + s_i$ for $i = 0, 1, \dots, 11$, and $7^{12} = 52n_{12} + 1$, from which (3) follows. Here are the lists described in the Theorem for a few values of N (see Table 1). I have included $N = 4, 8, 9, 12$ just to show that the Theorem applies in these cases, but Euler omits them for the simple reason that if you have the list for any N then you can trivially derive from it the list for Nk^2 for any k . For if p divides $a^2 + Nk^2b^2$ without dividing either a or Nk^2b then it divides $a^2 + N(kb)^2$ without dividing either a or Nkb , and on the other hand, if it divides $a^2 + Nb^2$ and if it does not divide k then it divides $(ka)^2 + Nk^2b^2$ without dividing either ka or Nk^2b .

A modern reader, after he sees the word **Theorem**, expects to find the word *Proof* soon thereafter. However, customs were different in Euler's day and his paper contains 59 theorems without a single proof. He told Goldbach in his letter that "I have no demonstration of it as yet," and the fact is that he never found a demonstration of it or even of a substantial portion of it. His "theorems" were based on nothing but empirical evidence.

In order to test the Theorem empirically one needs to be able to test, given a prime number p and a positive integer N not divisible by p , whether there exist integers a and b not divisible by p such that p divides $a^2 + Nb^2$. This at first looks impossible to test because it looks like one must test an infinite number of values of a and b . However, a moment's reflection shows that one need only test

Table 1.

N	list											
1	1											
2	1	3										
3	1	-5										
4	1	-3	5	-7								
5	1	3	7	9								
6	1	5	7	11								
7	1	-3	-5	9	11	-13						
8	1	3	-5	-7	9	11	-13	-15				
9	1	5	-7	-11	13	17						
10	1	-3	7	9	11	13	-17	19				
11	1	3	5	-7	9	-13	15	-17	-19	-21		
12	1	-5	7	-11	13	-17	19	-23				
13	1	-3	-5	7	9	11	15	17	19	-21	-23	25

values of a and b that are positive and less than p , because p divides $a^2 + Nb^2$ if and only if it divides $(a+p)^2 + Nb^2$ and the same holds for $a^2 + N(b+p)^2$, so multiples of p can be removed from a and b .

Using this observation, we can illustrate how one can test the Theorem, for example, for $N = 30$. Some numbers of the form $a^2 + 30b^2$ are

$$31, 34 = 2 \cdot 17, 39 = 3 \cdot 13, 46 = 2 \cdot 23, \\ 55 = 5 \cdot 11, 66 = 2 \cdot 3 \cdot 11, 79, \text{ and } 94 = 2 \cdot 47.$$

Thus the list must contain 31, 17, 13, 23, 11, 79 \equiv -41, 47, where \equiv indicates that 79 appears in the list as -41 when multiples of $4N = 120$ are removed to put the number between -60 and 60. More entries in the list can be found by using products of these. For example, $31 \cdot 17 = 527 \equiv 47$ is already in the list, $31 \cdot 13 = 403 \equiv 43$, $31 \cdot 23 = 713 \equiv 7$, $31 \cdot 11 = 341 \equiv 19$, $31 \cdot (-41) = -1271 \equiv 49$, and $31 \cdot 47 = 1457 \equiv 17$. A check shows that this assigns a sign to each positive integer less than 60 and relatively prime to 60 other than 1, 29, 37, 53, and 59. These are resolved by $17 \cdot 11 = 187 \equiv 53$, $13 \cdot 23 = 299 \equiv 59$, $23 \cdot 47 = 1081 \equiv 1$, $11 \cdot (-41) = 451 \equiv 29$, and $31 \cdot (-53) = -1643 \equiv 37$. Thus the list for $N = 30$ is 1, -7, 11, 13, 17, -19, 23, 29, 31, 37, -41, 43, 47, 49, -53, 59. For any prime p , the Theorem now gives a prediction as to whether p does or does not divide a number of the form $a^2 + 30b^2$ without dividing a or $30b$, and this prediction can be checked in a finite number of steps. For example, it predicts that 37 *does* divide a number of this form, and, indeed, $9^2 + 30 = 111 = 3 \cdot 37$. It predicts that 7 *does not* divide a number of this form, and, indeed, a check of the 36 numbers $a^2 + 30b^2$, $0 < a < 7$, $0 < b < 7$, shows that none of them is divisible by 7.

It is a long test to determine in this straightforward way whether a given p divides $a^2 + Nb^2$. The work can be greatly reduced by showing that if p divides any number of this type without dividing b then it divides a number of this type in which $b = 1$ and $0 < a < p/2$.¹ Thus in the case $p = 7$, $N = 30$, one need only check that 7 does not divide 31, 34, or 39 in order to conclude that the prediction of the Theorem is correct. Similarly, since 19 does not divide 31, 34, 39, 46, 55, 66, 79, 94, 111, the prediction for 19 is correct.

In a few hours one could verify in this way the prediction of the Theorem in thousands of cases for dozens of values of N . Because the Theorem is so simple and general and withstands these tests so easily, one readily becomes convinced that it is true. Certainly Euler was convinced, so much so that at times he seems to have forgotten that the Theorem was completely unproved.

For simplicity, the case of negative N , that is, of prime divisors $x^2 - Dy^2$ where $D > 0$, was omitted from the statement of the Theorem. It is easy to see that if D is a square then every prime p divides a number of this form. (For if $D = k^2$ then $x = k + p$ gives $x^2 - k^2 = p(2k + p)$, and p divides x only if it divides D .) However, if D is not a square then, as Euler already observed in his letter to Goldbach, a similar Theorem holds, except that instead of *never*

¹Here is the argument. Since p does not divide b and p is prime, 1 is the greatest common divisor of p and b . The Euclidean algorithm can therefore be used to write $1 = Ap + Bb$ for integers A and B . If p divides $a^2 + Nb^2$ then it also divides $B^2a^2 + NB^2b^2 = c^2 + N(1 - Ap)^2$ and therefore divides $c^2 + N$. Now $c = qp + r$ where the remainder r can be taken in the range $-p/2 < r < p/2$ and p divides $r^2 + N$, as was to be shown.

containing both x and $-x$ the list in these cases always contains both whenever it contains either.

Theorem (continued). *If N is negative and not of the form $-k^2$ then there is a list of integers s in the range $0 < s < |4N|$ and relatively prime to $4N$ such that (1), (2), and (3) hold (with $s < 4N$ changed to $s < |4N|$ in (3)). In this case (4) is replaced by*

(4') *Exactly half the positive integers less than $|2N|$ and relatively prime to $2N$ are in the list, and x is in the list if and only if $|4N| - x$ is in the list.*

For example, here are the lists for a few negative values of N written, as before, with $-x$ in place of $|4N| - x$. The first three are from Euler's letter (see Table 2).

Table 2.

N	list						
-2	± 1						
-3	± 1						
-5	± 1	± 9					
-6	± 1	± 5					
-7	± 1	± 3	± 9				
-10	± 1	± 3	± 9	± 13			
-11	± 1	± 5	± 7	± 9	± 19		
-13	± 1	± 3	± 9	± 17	± 23	± 25	

Actually, there is a simple relation between the lists for N and $-N$ which can be summarized by saying that a number x of the form $4n + 1$ is either in both lists or it is in neither. For example, for $N = 7$, the numbers 1, 9, -3 are in both lists and 5, 13, -11 are in neither. It is possible in this way to find either list once the other is known. The relation is simple to prove² and it was well known to Euler.

It would be difficult to exaggerate the importance of this Theorem in the history of number theory. The effort to prove it surely spurred much of Euler's own later work, and the other two great number theorists of the 18th century, Lagrange and Legendre, also worked on topics around and about the Theorem without penetrating the Theorem itself. Finally, the young Gauss found a proof in 1796, and

²If $p = 4n + 1$ then, by the case $N = 1$ of the Theorem (which is one of the few cases that Euler later succeeded in proving) p divides $y^2 + 1$ for some y . If p also divides $x^2 + N$ for some x not divisible by p —i.e., if p is in the list for N — then p divides $x^2y^2 + Ny^2 = (xy)^2 - N + N(y^2 + 1)$, which shows that p divides $(xy)^2 - N$ and therefore that p is in the list for $-N$. Since N is not assumed to be positive in this argument, the same argument shows that if p is in the list for $-N$ it is also in the list for N .

published two proofs in his great work, the *Disquisitiones Arithmeticae* in 1801. Gauss claimed to have discovered the Theorem on his own, but he would have needed to be in a cocoon in order not to have had some contact with work in this direction by Euler, Lagrange, and Legendre in the preceding half-century. I believe that Gauss was not being dishonest, but that he may have forgotten many subtle influences.

Gauss's formulation of the Theorem was very different from Euler's. For Euler, the basic question was whether, given N and p , the prime p divides a number of the form $x^2 + N$. It was noted above that if one can answer this question for N then one can easily deduce the answer for $-N$. A similar argument shows that if N is a product of two numbers $N = mn$ and if the question can be answered for each factor m, n then it can be answered for N . (This becomes clear when the question "Is p in the list for N ?" is restated "Is $-N$ a square mod p ?" as below. If the answer is known for m and $-n$ then it is known for $N = mn$ because a product is a square if and only if both factors are squares or neither factor is a square.) Thus it suffices to be able to answer the question for $N = 1$ and N a prime. The cases $N = 1$ and $N = 2$ were resolved by Euler and Lagrange, so the question was reduced to the case where N is an odd prime. Thus the problem is in essence to find the list in Euler's Theorem when $\pm N$ is an odd prime. One can find this list without testing a single prime divisor of $x^2 + N$ if one observes that the numbers common to the lists for N and $-N$, when N is prime, are precisely those numbers s , $-2N < s < 2N$, that can be written in the form $s = t^2 - 4Nk$ where t is a positive odd integer less than N . This is a simple consequence of the fact that squares are necessarily in the list.³

³To see this, note that if N is an odd prime then each list has $N - 1$ entries and half that many are common to the two lists. Therefore one need only show that all squares (reduced by subtracting multiples of $4N$ to put them between $-2N$ and $2N$) are in both lists, because this would account for all $(N - 1)/2$ common entries. For any of the $2N - 2$ nonzero odd integers x between $-2N$ and $2N$, multiplication by x and reduction by removing multiples of $4N$ is a one-to-one map of this set with $2N - 2$ elements to itself. For either of the two lists, if x is in the list then, by (3), multiplication by x carries elements of the list to elements of the list. Therefore, by counting, it carries elements not in the list to elements not in the list. In other words, if x is in the list and y is not then the reduction of xy is not in the list. Therefore multiplication by y carries elements of the list to elements not in the list. Since the list and its complement both have $N - 1$ elements, multiplication by y and reduction carries elements in the list one-to-one onto elements not in the list. By counting, then, it carries elements not in the list to elements of

For example, when $N = 11$, the numbers common to the lists are

$$1^2 = 1, 3^2 = 9, 5^2 \equiv 19, 7^2 \equiv 5, 9^2 \equiv -7;$$

thus the list for -11 is $\pm 1, \pm 9, \pm 19, \pm 5, \pm 7$, and the list for 11 is $1, 3, 5, -7, 9, -13, 15, -17, -19, -21$. When $N = 13$ the numbers in common are

$$1^2 = 1, 3^2 = 9, 5^2 = 25, \\ 7^2 \equiv -3, 9^2 \equiv -23, 11^2 \equiv 17$$

so the lists are $\pm 1, \pm 9, \pm 25, \pm 3, \pm 23, \pm 17$ and $1, -3, -5, 7, 9, 11, 15, 17, 19, -21, -23, 25$.

Gauss approached the subject from a different point of view, asking, for distinct odd primes p and q , whether q is a square mod p , that is, whether there is an integer x such that $x^2 - q$ is divisible by p . His “fundamental theorem,” now known as the **law of quadratic reciprocity** because it describes a reciprocal relationship between the questions “Is q a square mod p ?” and “Is p a square mod q ?” states:

If p is of the form $p = 4n + 1$ then q is a square mod p if and only if p is a square mod q .

If p is of the form $p = 4n - 1$ then q is a square mod p if and only if $-p$ is a square mod q .

This is easy to deduce from the Theorem above⁴, easy enough that it is not stretching matters very far to say that the law of quadratic reciprocity is a consequence of Euler’s theorems. However, for reasons to be explained in a moment, it is not in Euler’s interest to stretch matters at all.

The law of quadratic reciprocity is the crowning theorem of elementary number theory. One might almost say that it is the theorem with which elementary number theory ceases to be elementary. Gauss, who did not waste time with trivialities, was fascinated by this theorem, so simple to state and so difficult to prove, and he returned to it many times in his career, giving six different proofs of it.

Gauss also studied *higher* reciprocity laws, which deal, roughly speaking, with the prime divisors of $x^3 - N$ (cubic reciprocity), $x^4 - N$ (biquadratic reciprocity), etc. The study of higher reciprocity laws

the list. Therefore if y is not in the list, the reduction of y^2 is. Thus the reduction of y^2 is in the list whether or not y is.

⁴Here is the argument. If $p = 4n + 1$ and p is a square mod q , say $p - z^2$ is divisible by q , then $y = z$ or $z + q$ is odd and $p - y^2$ is divisible by both 4 and q . Therefore p is in the list for $N = -q$ (and also for $N = q$), which means that $x^2 - q$ is divisible by p for some x , that is, q is a square mod p . Conversely, if q is a square mod p then p is in the list for $N = -q$. Therefore, since $p = 4n + 1$, p is in both lists and $p = t^2 4qk$, which shows that p is a square mod q . The proof in the case $p = 4n - 1$ is the same with p replaced by $-p$.

was unquestionably the central question of 19th-century number theory, engaging the best efforts of Jacobi, Eisenstein, Kummer, Hilbert, and many others, and leading to the creation of algebraic number theory. Two developments in the subsequent history of the subject give further testimony to Euler’s genius and the importance of the theorems that he first announced to Goldbach.

First, a manuscript of Euler published in 1849 (he had died in 1783) showed that Gauss was not in fact the first to study higher reciprocity laws, but that Euler had already made some substantial progress on cubic reciprocity as early as 1749, and had not published his “theorems” in this field. For example, he stated the following conjecture:

Let p be a prime of the form $3n + 1$. Then 5 is a cube mod p if and only if the representation of p in the form $p = x^2 + 3y^2$ satisfies one of the 4 conditions (1) $y = 15m$, (2) $x = 5k$, $y = 3m$, (3) $x \pm y = 15m$, or (4) $2x \pm y = 15m$.

(Theorem III of the letter to Goldbach may or may not assert the existence of such a representation $p = x^2 + 3y^2$ whenever $p = 3n + 1$, depending on one’s interpretation of the phrase “contained in the form $3xx + yy$.” In any case, Euler later not only asserted the existence of such a representation, he proved it rigorously.) Euler gave no indication of how he arrived at this astounding set of conditions, and the fact that they are correct struck the editor of the relevant volume of his collected works (Vol. 5 of the first series) as “bordering on the incomprehensible.” However, the conjecture can be derived by applying the ideas described above to “imaginary primes” of the form $x + y\sqrt{-3}$ and finding the classes of imaginary primes mod $3 \cdot 5$ for which 5 is a cube.

The second testimony to Euler’s genius in the history of the subject is that later research showed that the “reciprocity law” approach to the subject was something of a blind alley. Hilbert in the 1890’s formulated the quadratic and higher laws in terms of a simple product formula which was generally regarded as a more natural way of describing the basic phenomenon, and in which there is no “reciprocity” but, rather, an explicit formula for determining (in the quadratic cases) which classes mod $4N$ contain prime divisors of $x^2 + Ny^2$. Later, in the 1920’s, the subject reached what is generally regarded as its culmination in the form of the **Artin Reciprocity Law**, which, again, has no element of “reciprocity” in it. Moreover, in the quadratic case, Artin’s Law is

almost exactly the Theorem we have stated, which was discovered by Euler nearly 200 years earlier.

References

1. L. Euler, Theoremata circa divisores numerorum in hac forma $paa \pm qbb$ contentorum, Enestrom 164, *Comm. Acad. Sci. Petrop.* 14 (1744/6), 1751, 151–181; also *Opera Omnia*, (1)2, 194–222.
2. P.-H. Fuss, ed., *Correspondance Mathématique et Physique*, Imp. Acad. Sci., St. Petersburg, 1843, vol. 1, pp. 144–153, reprint by Johnson Reprint Corp., New York and London, 1968.
3. A. P. Juskevic and E. Winter, eds., *Leonhard Euler und Christian Goldbach, Briefwechsel 1729-1764*, Akademie-Verlag, Berlin, 1965.

Afterword

For more information on Maclaurin, the reader can consult H. W. Turnbull, *Bicentenary of the Death of Colin Maclaurin* [10], which contains numerous articles about aspects of his work.

Florian Cajori expanded his arguments in the article in this section into a book, *A History of the Conceptions of Limits and Fluxions in Great Britain from Newton to Woodhouse* [3]. A more recent treatment of much of the same material is Niccolò Guicciardini's *The Development of Newtonian Calculus in Britain, 1700–1800* [8], and a good survey article on calculus in the first half of the eighteenth century is by H. J. M. Bos [1].

But the eighteenth century is the century of Euler. So to learn more about the mathematics of that century, it is essential to study the works of the Swiss genius. One good way to begin is with William Dunham's marvelous little book: *Euler: The Master of Us All* [4], which gives details of a number of Euler's mathematical gems. One can also read Euler's *Introduction to Analysis of the Infinite* [6], in an English translation by John Blanton. Although there is not yet a full-scale scientific biography of Euler, one good sketch of a biography is by Clifford Truesdell in the English translation of Euler's *Elements of Algebra* [9].

There are also histories of specific topics considered by Euler. For example, the history of analysis is well treated in Umberto Bottazzini, *The Higher Calculus: A History of Real and Complex Analysis from Euler to Weierstrass* [2] and Ivor Grattan-Guinness, *The Development of the Foundations of Mathematical Analysis from Euler to Riemann* [7]. Euler's number theory is a major topic in André Weil's *Number Theory: An Approach through History from Hammurapi to Legendre* [12], and details on some of Euler's work on algebra are found in B. L. van der Waerden's *A History of Algebra* [11]. Finally, Euler's complete works are still in the process of being published, the process having started in 1911. There are currently about 80 volumes available covering Euler's published works, with several more to come dealing with his letters and unpublished manuscripts [5].

References

1. Hendrik J. M. Bos, Differentials, higher-order differentials and the derivative in the Leibnizian calculus, *Archive for History of Exact Sciences* 14 (1974/75), 1–90.
2. Umberto Bottazzini, *The Higher Calculus: A History of Real and Complex Analysis from Euler to Weierstrass*, Springer, New York, 1986..
3. Florian Cajori, *A History of the Conceptions of Limits and Fluxions in Great Britain from Newton to Woodhouse*, Open Court, Chicago, 1919.
4. William Dunham, *Euler: The Master of Us All*, MAA, Washington, 1999.
5. Leonhard Euler, *Opera Omnia*, Societas Scientiarum Naturalium Helveticae, Leipzig, Berlin, and Zürich, 1911–.
6. Leonhard Euler, *Introduction to Analysis of the Infinite, Book I*, John D. Blanton.(trans.), Springer, New York, 1988.

7. Ivor Grattan-Guinness, *The Development of the Foundations of Mathematical Analysis from Euler to Riemann*, MIT Press, Cambridge, 1970.
8. Niccolò Guicciardini, *The Development of Newtonian Calculus in Britain, 1700–1800*, Cambridge University Press, 1989.
9. Clifford Truesdell, Leonard Euler: supreme geometer, in *Leonhard Euler, Elements of Algebra*, Rev. John Hewlett (trans.), Springer, New York, 1984, vii–xxxix.
10. H. W. Turnbull, *Bicentenary of the Death of Colin Maclaurin*, Aberdeen University Press, Aberdeen, 1951.
11. B. L. Van der Waerden, *A History of Algebra*, Springer, New York, 1985.
12. André Weil, *Number Theory: An Approach through History from Hammurapi to Legendre*, Birkhäuser, Boston, 1984.

Index

- Abacus, 148–151
Algebra, 42–45, 65–66, 143–147, 164–168, 171–172, 195, 288–291, 361–368
Algebra (L'Algebra), 164–167
Al-Kashi, 138–141
Almagest, 36–37, 55–56
Analytic Geometry, 189–197, 199–207, 244–247
Apollonius, 34–35, 51, 55, 203–204, 272, 348
Archimedes, 31–33
Aryabhata, 39, 124, 134
Astrolabe, 57
Astronomy, 256–259, 262–272, 343–344
- Babylonian mathematics, 5–26
Berkeley, G., 311, 313, 318, 325–327, 330
Bernoulli, J. and J., 186–187, 275–276, 332–334
Binomial series, 210–212, 252–254
Bombelli, R., 164–167
Brachistochrone, 186–7
Brahmagupta, 39, 134–135
- Calculating, 148–151
Calculus, 33, 77–80, 122–129, 179–187, 194–195, 218–234, 248, 252–254, 293, 310–321, 325–331, 369–374
Cardano, G., 153–163
Cauchy, A.-L., 225–226, 315–316, 355, 358
Chinese mathematics, 60–82
Chinese Remainder Theorem, 65
Chou pei suan ching, 62, 64
Complex numbers, 288–291
Conic sections, 30, 34, 272, 348
Conchoid, 205–206
Cotes, R., 238
Cubic equations, 153–163
Curve drawing, 199–207, 292–296
Cycloid, 183–187
- D'Alembert, J., 313, 315–317, 320, 340, 344, 354–355, 357, 362–363
Derivative, 218–227
Descartes, R., 184–185, 188–207, 244–248, 250, 292
Diez, J., 170–172
Differential equations, 223–224, 342
Differentials, 293, 369–374
Differentiation, 310–321
Diophantus, 38–39, 41–46, 51, 56
e, 346–352
Epistola posterior, 281, 285–286
Epistola prior, 253, 280, 284–285
Euclid's *Elements*, 30–31, 243–244
Euler, L., 223–224, 238, 317, 334, 336–345, 351–352, 354–359, 361–381
- Fermat, P., 122–123, 185–186, 218–220
Ferrari, L., 153–154, 159–162
Fibonacci, 143–147
Finger counting, 84–85
Fluxions, 310–321, 325–331
Function, 223–225, 354–356
Fundamental Theorem of Algebra, 361–368
- Gauss, C., 368, 379–381
Gaussian elimination, 63
Geography, 179–181
Geometry (La Géométrie), 188–197, 199–207, 244–248, 292
Greek mathematics, 27–59, 131–134
Gregory, J., 111, 114–116, 181, 208–216, 236–237, 255, 348
- Halley, E., 237, 252, 257, 261, 349
Harmonic series, 332–334
Hipparchus, 132
Hippias, 27–28
Hippocrates, 27
Hydroscope, 57
Hypatia, 47–58
- Ibn al-Haytham, 124–126, 136
Incas, 98–101
Indian mathematics, 39–40, 116–119, 126–129, 134–137
Institutiones Calculi Differentialis, 370–373
Interpolation formula, 209–210
Introductio in Analysin Infinitorum, 339–340, 369–374
Islamic mathematics, 123–126, 138–141
- Jyesthadeva, 116–118, 126–129, 135–136
- Lagrange, J. L., 224–225, 315–316, 321, 344, 355, 358
Leibniz, G.W., 111–114, 186, 221–222, 288–295, 350, 369–370
Leonardo of Pisa, 143–147
Letters to a German Princess, 340–341

- Liber Quadratorum*, 143–147
 Linear perspective, 303–308
 Liu Hui, 69–80
Lo shu, 60–61
 Logarithms, 235–238, 347–349

 Maclaurin, C., 209, 224, 310–321, 328–331
 Mayan mathematics, 94–96, 101–103
 Mercator, G., 179–180
 Mercator, N., 113, 235–236, 252, 349
 Mesopotamian mathematics, 5–26

 New World, 169–170
 Newton, I., 122, 209–211, 221–223, 231–234, 237, 240–287, 314, 327–328
 Newton's method, 279–286
 Nilakantha, K.G., 111–112, 116–119, 126, 135
Nine Chapters on the Mathematical Art, 63–65, 69–80
 North American Indians, 83–93
 Number systems, 88–91, 94–96, 148–150
 Number theory, 38–39, 340–341, 375–381

 Optics, 254–256, 341–342

 Pappus, 37
 Parabola, 293–295
 Partial differential calculus, 354–359
 Pascal, B., 186
 Pascal's triangle, 66
 Perspective, 303–308
 Pi, 75–76, 111–119,
 Plato's Academy, 29
Plimpton 322, 7–12, 14–25
 Polar coordinates, 274–277
Principia mathematica, 256–259, 262–272
 Projective geometry, 37, 303–308

 Ptolemy, 20, 36–37, 51
 Pythagoras, 27
 Pythagorean triples, 10–12, 15–17

 Quadratic reciprocity, 375–381
 Quipu, 98–101

 Reciprocals, 12, 21–24
 Roberval, G.P., 122–123, 183–185, 228–231
 Robins, B., 327–330

 Schooten, F. van, 248–249
Sea Island Mathematical Manual, 74–75
 Secant, 179–181
 St Vincent, G., 347
 Square roots, 64–65, 72–73
Sumario Compendioso, 169–172

 Tangents, 228–234
 Tartaglia N., 153–163
 Taylor, B., 303–309
 Taylor series, 111–119, 208–209, 223–224, 231–238
 Thales, 27
 Theon, 47, 52, 55–56, 58
Treatise of Fluxions, 310–321
 Trigonometry, 18–20, 35–37, 131–141

Vera Quadratura, 212–216
 Volume of a pyramid, 76–78
 Volume of a sphere, 79–80

 Wallis, J., 113, 249–250, 253, 349–350
 Weierstrass, K., 226
 Woodhouse, R., 330
 Wright, E., 180–181

About the Editors

Marlow Anderson is a professor of mathematics at The Colorado College, in Colorado Springs; he has been a member of the mathematics department there since 1982. He was born in Seattle, and received his undergraduate degree from Whitman College. He studied partially ordered algebra at the University of Kansas and received his PhD in 1978. He has written over 20 research papers. In addition, he is co-author of a book on lattice-ordered groups, and also an undergraduate textbook on abstract algebra.

Victor Katz is currently Professor of Mathematics at the University of the District of Columbia. He has long been interested in the history of mathematics and its use in teaching. The first edition of his textbook: *A History of Mathematics: An Introduction* was published in 1993, with a second edition in 1998 and a shorter version to appear in 2004. He has directed three major NSF-supported and MAA-administered grant projects dealing with the history of mathematics, collectively titled the *Institute in the History of Mathematics and Its Use in Teaching* (IHMT). Under these projects, over a hundred college faculty (and thirty-five high school teachers) studied the history of mathematics, including how to teach courses in the subject and how to use it in teaching mathematics courses. In the third of the projects, the *Historical Modules Project*, eleven modules were developed for teaching topics in the secondary mathematics curriculum via the use of history. These are available now on a CD.

Robin Wilson is currently Head of the Pure Mathematics Department at the Open University, U.K., and Fellow in Mathematics at Keble College, Oxford University. He was Visiting Professor in the History of Mathematics at Gresham College, London, in 2001–02 and is a frequent visiting professor at Colorado College. He has written and edited about 25 books, in topics ranging from graph theory and combinatorics, via philately and the Gilbert & Sullivan operas, to the history of mathematics. In 1975 he was awarded a Lester Ford award by the MAA for “outstanding expository writing.” He is well known for his bright clothes and atrocious puns.