Michael A. Dopita & Ralph S. Sutherland

Diffuse Matter in the Universe, Subtitle, Edition

Physics – Monograph
(Editorial W. Beiglböck)

January 9, 2001

'. . . Light is a confused aggregate of Rays indued with all sorts of Colours, as they are promiscuously darted from the various parts of luminous bodies. A naturalist would scarce expect to see ye science of those colours become mathematicall, and yet I dare affirm that there is as much certainty in it as in any other part of Opticks.'

Sir Isaac Newton, in a letter to the Royal Society, 1672

Preface

Here come the golden words

place(s),   *Firstname Surname*
month year   *Firstname Surname*

Contents

1. What is Diffuse Matter?

"Now entertain conjecture of a time
When creeping murmur and the pouring dark
Fills the wide vessel of the Universe"
— Shakespeare (Henry IV, Act4)

Nowhere in the universe can we find a perfect vacuum. Around ordinary stars, a hot magnetized plasma seethes and bubbles outward in a thermally-powered wind. Within the disks of spiral galaxies, dusty clouds of molecular gas continually coalesce and collapse under their self gravity to form new stars. The light from these new-born stars heats and ionizes their placental cloud before finally dispersing it back into the galactic disk, ready to repeat the cycle. Wherever there are stars, some reach the end of their lives and explode as supernovae, hurling out the gas that transformed in their thermonuclear furnaces to heavy elements such as iron. The shocks they produce rumble their way through space, heating the interstellar gas anew. Some is thrown up and far away from the galactic plane, while other parts are crushed into dense sheets and filaments which shine briefly as their shock energy is radiated away. In clusters of galaxies, the gaseous shells ejected by old stars collide one with another, heating the gas so that it glows softly in X–rays, cooling over billions of years before finally falling back into the bright galaxy cores to feed the massive monster black holes that lurk at their centres. In intergalactic space jets of relativistic plasma shot from the cores of active galaxies emit radio waves as charged particles circle and shed their energy in the magnetic fields. Even in the vast reaches of space between the clusters of galaxies, hot plasma can still be found, even though it is so tenuous that it can never cool again, and a hydrogen nucleus could travel a distance equal to the width of our galaxy before encountering another of its kind.

This then is the broad canvas of diffuse matter in the universe, displaying a rich range of thermal plasma phenomena and covering a remarkable variety of conditions and chemical compositions. We work to understand the physics of such plasmas because, by gaining insight into their physics, we can hope to understand and interpret the observed phenomena, measure their physical parameters and determine their chemical composition. Since all the stars in all the galaxies have been ultimately formed from this gas, our study provides insight into the structure and evolution of the universe we live in.

Fig. 1.1. Turbulent ionised plasmas filling a nearby region of space around the active star-forming region of the Orion Nebula.

What sets the physics of cosmic gas clouds apart from the molecular gas found in the atmosphere of our planet, or the hot ionized plasmas found in the interiors of stars? The key to this question lies in the densities that characterize the interstellar and intergalactic plasmas. We generally refer to the density of the gas in units of atoms per cubic centimeter. The mean density of the gas in the plane of our galaxy at the solar radius is about $0.3\,\mathrm{cm}^{-3}$. Inside a molecular cloud the density is much higher, say, $10^6\,\mathrm{cm}^{-3}$. Nonetheless, this is completely negligible compared with the density encountered in the earth's atmosphere ($\sim 2 \times 10^{18}\,\mathrm{cm}^{-3}$). In our atmosphere, an atom collides with another in a time–scale of a few nanoseconds. In our molecular cloud the collision time–scale is extended to several days. However, internal atomic or molecular time–scales range from around a nanosecond to several millions of years for the neutral hydrogen $21\,\mathrm{cm}$ emission line. Thus each atom that suffers a collision in the earth's atmosphere will suffer another encounter before it has a chance to adjust internally, through atomic processes, from the previous collision. For most processes, the atom in our molecular cloud is free to radiate any excess internal energy it has picked up as a result of the collision, and return to its ground state.

As a result, the atoms of our diffuse astrophysical plasmas are usually sitting in their ground state, while the atoms in the earth's atmosphere are in a dynamical balance with collisional processes, every possible atomic state being fed by collisions as fast as it is being depleted by other collisions. This condition is known as *Local Thermodynamic Equilibrium* (LTE). In this case, at a temperature T, the number density of atoms or molecules in any excited state, j, compared with the number in the ground state ($j = 1$), is given by the *Boltzmann equilibrium:*

$$\frac{N_j}{N_1} = \frac{g_j}{g_1} \exp[\frac{-\Delta E_m}{kT}] \qquad (1.1)$$

here, k is the Boltzmann Constant ($k = 1.380622 \times 10^{-16}\,\mathrm{erg\,K}^{-1}$), ΔE_m is the energy difference between the ground and the excited state; and g_j is the statistical weight of the excited state (if the excited level has a total angular momentum quantum number J, then $g_j = 2J + 1$). In addition, the electrons, and each species of atom or molecule in the gas will have their energy distributed according to the *Maxwell Distribution:*

$$n(E) = \frac{2N}{\pi^{1/2}(kT)^{3/2}} E^{1/2} \exp[\frac{-E}{kT}]dE \qquad (1.2)$$

The low densities in the diffuse plasmas ensure that Boltzmann equilibrium is rarely, if ever, achieved, although in most cases, collisions between like species are sufficiently frequent to set up the Maxwell distribution. However, we will encounter a few examples of gas which has passed recently through a shock so that its ions have not had time to achieve the Maxwell distribution, and in some cases it is even difficult to define the energy distribution

of the electrons. A plasma in which the Boltzmann equilibrium is not a good approximation is said to be in a *non–LTE (or NLTE)* condition. The study of the physics of the interstellar medium is therefore, in large part, the study of highly NLTE plasmas.

Historically, diffuse astrophysical plasmas have been divided into broad environmental divisions or domains, of which the most familiar and the best studied is the medium between the stars in our galaxy. This medium, and the gas in between stars in other galaxies is generally referred to as the interstellar medium (ISM). It provides in its dense molecular phase the cradle and the birthplace of stars. In turn, it derives its complex phase structure and energy balance from the input of energy derived from nuclear burning occurring within these stars in the form of photons, stellar winds and outflows or stellar explosions.

Within the ISM itself, the details of these interactions has provided a potent testing ground for theory, and has provided a good deal of insight into the evolution of stars, and the chemical and structural evolution of galaxies. The dense molecular clouds have given us an understanding of molecular chemistry, interstellar dust physics, the diffusion of magnetic fields, and a detailed insight into the gravitational instabilities which lead to the formation of stars. Star formation regions embedded within them have led to an understanding of how newly forming stars shed the angular momentum of their parent cloud through the formation of collimated outflows and jets, and the interaction of these jets with the clouds has enabled us to understand shock physics and chemistry. Radio observations of molecular lines enable us to study physical conditions in star–forming regions, and tell us about the isotope ratios of various elements, which are key data in understanding the sites of production of the heavier elements.

The photoionized regions around massive young stars provide a means of probing the chemical composition of the atomic gas both in our Galaxy, and in distant galaxies. In particular, they provide us with estimates of the primordial helium abundance – a key parameter in cosmological models. We can also study the bubbles formed by the powerful radiation-pressure driven winds of their central stars.

The so–called planetary nebula shells which have been ejected and photoionized by dying solar–like stars provide insight into the chemical processing and dredge–up which has occurred in their atmospheres, and give us an observational test of stellar evolution in low mass stars, while the nova shells ejected from the surfaces of White Dwarf stars enable us to study explosive nuclear processing under electron–degenerate conditions.

Finally, the material ejected in supernova explosions gives us a sample of the end–products of nucleosynthesis in stars, and the properties of the shock waves driven into the surrounding medium measures the kinetic energy input into the ISM by these explosions, while the properties of the radio synchrotron

spectrum generated in the shell provides insight into particle acceleration mechanisms and the origin of cosmic rays.

With each generation of stars, some of the ISM is lost forever in the dying embers of stars – the White Dwarfs, in neutron stars formed during supernova explosions, or in Black holes formed in the collapse of the cores of massive stars. In addition, some matter is effectively lost in low–mass stars which frugally burn their nuclear fuel over time–scales much longer than the age of the universe, while yet another part is stored for long periods within normal stars like our own sun. In the solar neighborhood, the ISM now accounts for only about 15% of the total baryonic mass, and this figure is typical of spiral galaxies.

However important the gas and dust in galaxies may be, we must not forget other parts of the cosmos where we find components that are not moulded and controlled by stars alone, and indeed may not even located between stars as the word interstellar implies. In the cores of many galaxies lurk massive black holes which when fed by matter subject their environs to extremes of ionization or temperature. Here we find rings of gas which are irradiated by X–rays, massive outflows of material, or highly relativistic jets of gas shot out into intergalactic space. These plasmas could collectively be regarded as the Active Galactic Medium (AGM). Developing an understanding of the properties of the AGM is a cornerstone of research in active galaxies.

Finally, on the largest of scales we have the intergalactic medium (IGM) or the hot gas found within whole clusters of galaxies, the intra-cluster medium (ICM). This material is detected by means of the X-rays it produces, by the effect it has upon the propagation of relativistic jets, by faraday rotation and depolarization of distant radio sources, by evidence of ram-pressure stripping of matter in galaxies, or through the absorption it can produce in the light of distant galaxies. Since the IGM is the most difficult to observe, it is also the least well studied of the diffuse astrophysical plasmas.

1.1 Phases

Diffuse matter in the universe is found over an extraordinary range of scales; from structures smaller than the size of the solar system ($\sim 10^{15}$ cm), up to regions encompassing whole clusters of galaxies ($\sim 10^{24}$ cm). Fortunately, the characteristic densities decrease as the scale size increases (otherwise the mass of the diffuse medium would be infinite!). A consequence of this enormous dynamic range in parameter space is that the range of possible phenomena is very rich. However, provided that characteristic time–scales remain appreciably shorter than the characteristic time–scale for the evolution of the universe ($\sim 10^{10}$ yr), it is remarkable how often phenomena occurring on small scales find physical analogues on much larger scales. As an example, we might cite the bipolar outflows from very young stars which have characteristic scales 10^5 yr and 10^{17} cm, but which have very similar physics to the bipolar jets

in active galactic nuclei (AGN), with characteristic scales of order 10^8 yr and 10^{22} cm.

The diffuse medium has always represented an important component of the baryonic mass in the universe. Shortly after the recombination epoch in the early expansion of the universe, diffuse matter reigned supreme, accounting for all the baryonic mass, with the exception of a few primordial Black Holes. However, gravity soon took over, and the gas clouds started coagulating and collapsing. It should always be remembered that gas is sticky stuff. When two gas clouds collide, their energy of motion is turned into heat through shocks, which is then radiated away into space. Thus cloud–cloud collisions can be considered as almost completely inelastic, conserving momentum, but losing all their kinetic energy. Shocks are also highly compressive, and aid the development of a cold dense phase in the ISM which in turn favors star formation when the densities become high enough. Early on in the universe, these processes led to the formation and the evolution of galaxies, and today the ISM in galaxies is kept in a dynamic, self-regulating equilibrium determined by the rate of star formation, balanced against the energy input these stars put back into the interstellar medium. In galaxies, the ISM forms a multi–phase structure in response to this feedback, and develops a heirachical fractal spatial structure.

The multi–phase structure, discussed in detail in Chapter (14), develops as a consequence of the fact that a stable balance of heating and cooling at a given pressure can often be achieved at more than one temperature. Various names have been given to the most common phases of the ISM in galaxies. The molecular medium (MM), the cold neutral medium (CNM) and the warm neutral medium (WNM) are three such phases of the atomic gas in the ISM. In similar fashion, we may also find components due to a warm ionized medium (WIM) and a hot ionized medium (HIM). None of these components should regarded a static in time or space, and matter is constantly in flux between them.

Dying stars constantly feed matter back into the ISM, which has been transformed into heavier elements of one kind or another (often labelled, in cavalier fashion, "metals" by astronomers). These processes of chemical evolution are very clearly described by Pagel (1997). Much of the non-volatile fraction of heavy elements finds its way eventually into interstellar dust grains, which are important constituents of the ISM, absorbing and polarizing the light from distant stars, coupling gas and magnetic fields through photoelectric or collisional charging, playing an important role in the total energy balance of the ISM, and providing on their surfaces sites for chemical reactions which allow complex molecules to form.

If the plasma is hot, its lifetime in the hot phase depends on its heat content or internal energy, and how fast it can radiate this heat away. The rate of radiation is a complex function of temperature and of density. Thus, we can arrange the plasmas we meet in the ISM according to their characteristic scale

size, density and temperature. This has been done in Figure (1.2). The realm of the molecular clouds and the H I regions is the realm of astrochemistry, while all the other classes of object are predominantly regions of ionized plasmas of one kind or another. Although an approximate division between LTE and NLTE plasmas is indicated, this transition is rather fuzzy, and depends on the particular ion or atom considered. Although, as we will see, fully NLTE plasmas are reasonably straightforward to understand, and so are fully LTE plasmas, the region close to the transition zone between the two cases is computationally awful, since all levels of all atoms have to be considered together along with their associated radiative transfer problems. This is certainly not a physical regime that can be left to the student, but requires the full capability of modern supercomputers and unlimited atomic data. In this book, we will consider only thermal processes in low-density plasmas, leaving these transitional plasmas, relativistic plasmas and their emission processes to other texts.

1.2 Observability

It might seem rather trivial to remark that, in order for a given phase of the ISM to be detectable in its own emission lines or continuum, it must emit enough photons in any wavelength band to which our earth-bound or space-borne telescopes are sensitive. In practice, as we will see below, these emission processes are mainly governed by binary collisions between electrons and the ion responsible for the line of continuum emission, or by collisions between electrons. As a consequence, the local emissivity (measured in $\mathrm{ergs\,cm^{-2}s^{-1}sr^{-1}}$) varies as the square of the local density. On the other hand, the surface brightness of an object is governed by whether this local emissivity, integrated along the line of sight, is sufficient to be detectable against the background (zodiacal emission, airglow, telescope emissivity, instrumental noise, or whatever). Therefore, a very useful parameter characterizing ISM sources is the so-called *Emission Measure* defined as:

$$EM = \int n_e^2 dl \qquad (1.3)$$

In many cases, since we do not really know the run of density along the line of sight, this integral is approximated by the mean density and pathlength through the region of interest, $EM = \langle n_e \rangle^2 l$. Since astronomers like to define their densities in units of number densities ($\mathrm{cm^{-3}}$)), but their distances in units of parsecs ($1\,\mathrm{pc} = 3.0856 \times 10^{18}\,\mathrm{cm} = 3.2615$ light years), the emission measure is generally given in the somewhat ridiculous units of $\mathrm{pc\,cm^{-6}}$.

As illustrative examples, consider these cases of ionized hydrogen plasmas. First, the case of a Nova shell which has been ejected from the surface of a White Dwarf at a typical velocity of $1000\,\mathrm{km\,s^{-1}}$. After typically 100 days,
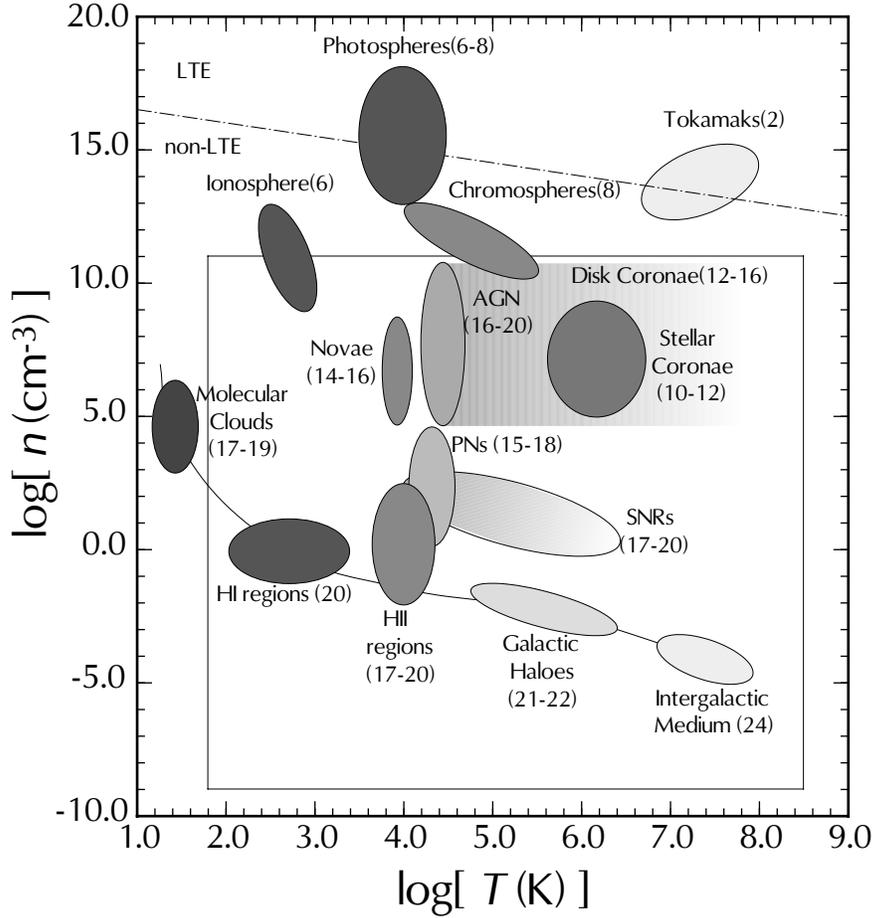
Fig. 1.2. Densities and characteristic sizes of diffuse astrophysical plasmas in the universe. For each class of objects, the characteristic size in log(cm) is given. The approximate boundary between plasmas in LTE and non-LTE plasmas is marked as a dash-dot line. Current nebular modelling is valid within the approximate boundaries of the box-shaped region. The thin solid curve connects the dominant phases of galactic and intergalactic diffuse media.

the shell has reached the nebular phase, that is to say it has become optically thin to the passage of radiation, and at that time it has a typical density of $10^7 \, \mathrm{cm}^{-3}$. In this case the EM $\sim 10^9 \, \mathrm{pc} \, \mathrm{cm}^{-6}$. As a second example, take a typical Planetary Nebula. This is the ionized envelope of a dying star as it transits from a red giant to a white dwarf star. Typically, the envelope will have been expanding at $10–30 \, \mathrm{km} \, \mathrm{s}^{-1}$ for a few thousand years, so that it is about 0.1 pc across. At this time it has a density of $10^4 \, \mathrm{cm}^{-3}$, and the

EM $\sim 10^7 \, \mathrm{pc\,cm^{-6}}$. Still fainter are H II regions, which are typically 10–100 pc across and are ionized by the UV light of massive, hot, and young stars a few million years old. In this case the density is as low as $10 \, \mathrm{cm^{-3}}$, and so the EM is only $10^{3-4} \, \mathrm{pc\,cm^{-6}}$. Such nebulae are still easy to detect with modern telescopes, although some faint lines which are important to establish the density, temperature or abundances may be difficult to observe. Finally, consider the case of the diffuse galactic ISM which pervades the disks of spiral galaxies. Here the densities are as low as $0.1 \, \mathrm{cm^{-3}}$, while the medium is limited by the scale height of the gas in the galactic plane, typically 150 pc. In this case the EM is only $\sim 1 \, \mathrm{pc\,cm^{-6}}$, and specialist instruments are needed simply to detect it, let alone measure it accurately enough for analysis.

Even when the plasma is too faint to be seen by its own emission, it may still be detected by the absorption it produces in a background source of continuum emission. This is because resonance transitions to higher states are excited by the continuum light in the beam, but the atom re-radiates this light in all directions when it returns to the ground state. Thus effectively the light has been scattered out of the line of sight. The absorptions are proportional to the column density $\int n_e dl$, and the cross–sections for absorption are large, comparable with the Bohr radius. Thus, species which have column densities as low as $10^{12-13} \, \mathrm{cm^{-2}}$ along the line of sight may be detected with high dispersion spectrographs on large telescopes. Such absorption techniques are the only means whereby the hot, highly ionized, and very tenuous gas in our Galactic halo can be detected. This gas has typical densities of $0.001 \, \mathrm{cm^{-3}}$, and columns of $\sim 3 \, \mathrm{kpc}$ ($10^{22} \, \mathrm{cm}$), so its emission measure would only be of order $10^{-2} \, \mathrm{pc\,cm^{-6}}$, and therefore totally undetectable in its own emission.

2. Line Emission Processes

"All Science is either physics or stamp collecting"
— Ernest Rutherford

In this chapter we will review the sometimes arcane, frequently confusing, notations that are used to classify atomic and molecular transitions and spectra. A basic understanding of these is essential, since it is through the atomic and molecular lines that we ultimately derive our understanding about the physical conditions in, and chemical abundances of, diffuse astrophysical plasmas. In what follows, we will assume that the reader is familiar with the basic concepts of quantum mechanics, usually developed during the second year of a physics major. If not, the reader is referred to the highly readable account of this and other topics in modern physics which is to be found in the book by Rohlf (1994), and a lucid introduction to molecular spectroscopy is to be found in Atkins (1983); *see* notes on this chapter.

2.1 Atomic Spectra

2.1.1 Resonance Lines

Let us first consider a simple two–level atom, as in Figure (2.1).

If the transition shown is a resonance line, it arises from a normal electronic dipole radiative transitions. Such a line is called a permitted line. This means that the transitions follow the standard selection rules of quantum mechanics which require that:

- Only one electron is involved in the transition.
- The initial and final states have different parities.

In addition, selection rules are imposed by the requirement of conservation of angular momentum:

- The change in the magnetic quantum number has to be $\Delta m_l = 0, \pm 1$.
- The intrinsic angular momentum quantum number of the electron, m_s, does not change; $\Delta m_s = 0$.
- Since the photon carries one unit of angular momentum, the electron orbital angular momentum must change by one unit, $\Delta l = \pm 1$.
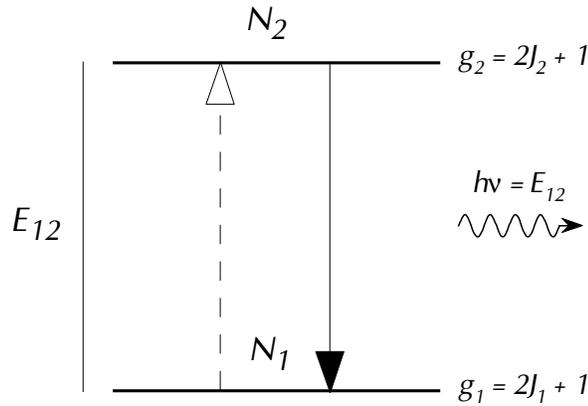
Fig. 2.1. An idealised two–level atom.

These imply that

- The change in the electron total angular momentum quantum number must be $\Delta j = 0, \pm 1$, with the restriction that the transition $j = 0 \rightarrow j = 0$ is forbidden (see Section 2.1.4).

Consider two states of an atom with total angular momentum quantum numbers J_2 and J_1. Since the total angular momentum includes the half–integral electron spin, the possible values of J are 0, 1/2, 1, 3/2, 2 *etc*. The statistical weight of any energy level, m, is given by $g_m = (2J_m + 1)$. The strength of a transition is measured by the probability that an excited atom will emit a photon in unit time. Thus, if there are $N(0)$ atoms in their excited state at time $t = 0$, and the transition probability is A (s^{-1}), then $N(t) = N(0)\exp[-At]$. The transition probability for line emission between these two states is proportional to ν^3 and is given by the Einstein relation:

$$A_{21} = \left(\frac{64\pi^4}{3hc^3}\right) g_2^{-1} \nu_{12}^3 S_{21} \text{ s}^{-1}, \qquad (2.1)$$

where S_{21} is the *Line Strength*, given in terms of the *electric dipole matrix element* by:

$$\mu_{21}^2 = S_{21} g_2^{-1} = e^2 |\langle \Psi_1 |r| \Psi_2 \rangle|^2 \sim e^2 n^4 a_0^2 / Z^2. \qquad (2.2)$$

Here $n^2 a_0/Z$ is the radius of the electron orbit in terms of the Bohr radius. Note that the electric dipole matrix element is symmetric between the states

involved ($\mu_{21} = \mu_{12}$), and may be interpreted physically as the strength of the electric dipole in the system during the transition. Note also that equation (2.1) implies a very strong dependence of the transition probability upon the frequency. Thus, if there are two paths for radiative decay, the branching ratio will always strongly favour the shorter wavelength transition. Resonant transitions are characterised by large transition probabilities; typically of order $\sim 10^{8-9}\,\mathrm{s}^{-1}$.

When an atom in its ground state is placed in an electromagnetic field with energy density[1] $U\left(\nu_{12}\right)$ (energy per unit volume; $\mathrm{erg\,cm}^{-3}$) at the frequency of the transition, the probability per atom that it will absorb one of these photons, and thus be excited into the upper level is $B_{12}U\left(\nu_{12}\right)$ where B_{12} is the *Einstein coefficient for absorption*, given in terms of the transition probability:

$$A_{12} = \left(\frac{8\pi h}{c^3}\right)\nu_{12}^3 B_{12}\ \mathrm{s}^{-1}, \qquad (2.3)$$

It follows from symmetry that,

$$g_1 B_{12} = g_2 B_{21} \qquad (2.4)$$

where for an atom in its excited state, the probablility per atom that it will absorb one of these photons, and thus be de–excited into the lower level is $B_{21}U\left(\nu_{12}\right)$ and B_{21} is called the *Einstein coefficient for stimulated emission*. Thus the relative importance of absorption to stimulated emission simply depends on the ratio of atoms in the ground state compared with atoms in the excited state. This ratio is closer to unity for transitions in the IR than for optically–emitting transitions. Thus the relative importance of stimulated emission is more important for IR transitions. In some circumstances, collisional pumping or radiative pumping may cause the upper level can become overpopulated, producing a population inversion. This allows the possibility of astrophysical masering, which will be discussed below.

Frequently we use the concept of *Absorption Oscillator Strength*, f_{12}. This treats the transition as if it were a bound harmonic oscillator, and asks how effective is the transition (in terms of equivalent numbers of classical electrons of cross–section, πr_0^2 ; where r_0 is the classical radius of the electron, $\mathbf{e}^2/m_e c^2$) in absorbing radiation from the radiation field at frequency ν_{12}). It is related to the transition probablity by:

$$A_{12} = \left(\frac{8\pi^2 \mathbf{e}^2}{m_e c^3}\right)\nu_{12}^2 f_{12} \qquad (2.5)$$

This is not symmetric between the initial and final states, but the gf value of the transition is:

---

[1] The energy density $U(\nu)$ should be distinguished from the brightness, or specific intensity, $B(\nu)$, which is the energy per unit area per unit solid angle and per unit time; $B = cU/4\pi$

$$gf = g_1 f_{12} = g_2 f_{21} \qquad\qquad (2.6)$$

here, f_{21} is called the *Emission oscillator strength*. The oscillator strengths for all transitions in the atom obey the Reiche–Thomas–Kuhn sum rule, namely, for an atom with Z optically–active electrons:

$$\sum_{n=1}^{\infty} f_{mn} = Z \qquad\qquad (2.7)$$

This follows from the definition of oscillator strength in terms of the equivalent numbers of classical electrons; clearly the total cannot exceed the number of electrons that are available to absorb.

In nebulae generally, thanks to the NLTE conditions which prevail, most atoms are sitting in their ground state, and the excited states have only a very low population. Therefore, transitions between excited states appear strong only as a result of recombination and cascade down towards the ground level. Such recombination lines which are commonly observed are the Balmer Series, Hα, Hβ *etc.* or recombination lines of He I and He II. These are seen in preference to the recombination lines of heavier elements thanks to the very large relative abundances of hydrogen and helium in the ISM.

2.1.2 Pure Recombination Lines

Hydrogenic ions are ions with a single electron bound to a nucleus of charge Z. This simple atomic system is one of the best studied and understood of all atomic systems. Because it consists, in the case of hydrogen itself, of just two distinguishable particles, the quantum mechanical wavefunction description is quite soluble. Even though strictly speaking heavier *hydrogenic* ions, such as C VI, have many particles because of complex nuclei, on the atomic scale the nucleus can be considers as a single particle to a good approximation. Historicly, observations of the hydrogen spectrum in astrophysical plasmas have provided stringent tests of the quantum mechanical model, and predictions of high precision have been verified.

In the limit of low density, the hydrogen spectrum is dominated by a process called recombination cascades. Electrons from the surrounding plasma have combined with a bare hydrogen nucleus to from a hydrogen atom in an excited state. The details of this   *recombination* process are deferred until chapter 6, here we will concentrate on the processes of line emission after recombination as the recombined electron cascades down through the many possible hydrogen energy levels to the ground state through resonance lines. The principle difference between this emission line process and the two level resonance line model shown earlier is the nearly *infinite* number of levels involved and the fact that electrons are appearing from 'above'. If we just consider the energy levels and their probability wavefunctions, we can calculate the essential parameters of the cascade process, the transition probabilities,

to great precision. Precision is necessary because the final spectrum depends on many probabilities all multiplied together and errors could accumulate unless they are very small to begin with.

It has been known for a long time that the hydrogen system with two particles orbiting each other with masses m_p for the proton and m_e can be treated equivalently as a reduced mass system. We have an atom with two particles of mass m_1 and m_2 separated by a distance r spinning about the center of mass located at distance r_1 and r_2 from each atom, respectively. The moment of inertia of the atom is given by:

$$I = m_1 r_1^2 + m_2 r_2^2, \tag{2.8}$$

which can be expressed in terms of the reduced mass of the system as

$$I = \frac{m_1 m_2}{m_1 + m_2} r^2 = \mu r^2 \tag{2.9}$$

where μ is the reduced mass at the radius r from the centre of rotation

The reduced mass hydrogen atom system can be described in terms of spherical wavefunctions Ψ that satisfy the of the time–independent Schrödinger equation:

$$H\Psi = E\Psi, \tag{2.10}$$

where H is the Hamiltonian operator and E is the energy.

Wavefunctions that satisfy the Schrödinger equation for this system, using spherical coordinates can be separated into angular and radial functions.

$$\Psi(r, \theta, \phi) = R(r)\Theta(\theta)\Phi(\phi). \tag{2.11}$$

The functions $\Theta(\theta)$ and $\Phi(\phi)$ are shown in figure **??** in the next section.

The eigenvalues or energy levels associated with each of the solution wavefunctions are quantised in three variables; n, l, m and result in a series of discrete, degenerate, energy levels determined by n alone of

$$E = -\frac{m_e \mathbf{e}^4 Z^2}{2\hbar^2 n^2}. \tag{2.12}$$

These energy levels are typically illustrated as shown in figure 2.2. Vertically we have energy with levels determined by the principle quantum number n, and horizontally the levels are spread out according to the angular momentum number l with no vertical displacement. In more complex atoms with many electrons, the repulsion between electrons results in vertical differences in the l levels. Here there is no repulsion and all the l levels are degenerate. The m levels are also degenerate because under spherical symmetry there is no prefered direction to put the axis around which m operates. The levels due to m are not shown, apart from needing a 3D figure the m levels are not distinguishable unless a strong external magnetic field is applied to the atom.
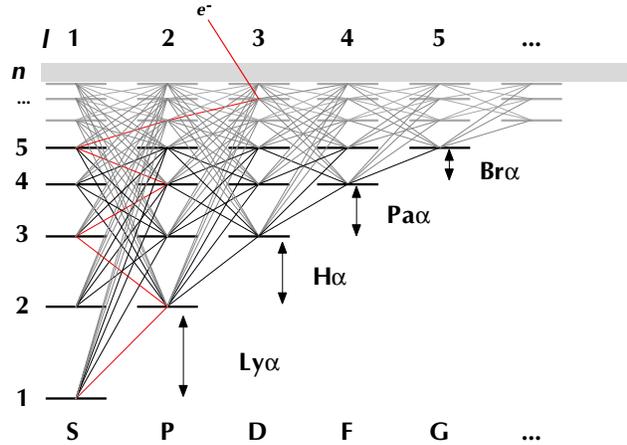
Fig. 2.2. Figure Hydrogen n, l, Levels

This can occur in extreme conditions such as compact star atmosphere, but it does not occur to any extent in diffuse plasmas.

The wavefunctions associated with each energy level have analytical forms. The angular parts $\Theta(\theta)\Phi(\phi)$ from the spherical harmonics, analagous to a vibrating string, but on the surface of a sphere. These functions are smooth and wrap around the sphere exactly as you might expect. They are described by a polynomial series in θ and ϕ known as the *Legendre Polynomials*. The radial functions also have analytical form

As with the resonance, intercombination and forbidden transitions already covered, the transitions between all these levels obey selection rules. The dominant one is the usual rule that the unit angular momentum of the emitted photon requires that a corresponding unit change of angular momentum of $\Delta l = \pm 1$ must take place. This rule is particularly strict in the hydrogen case and except for one special case, only resonant, permitted transition are observed. In the figure the allowed transitions are shown and form a characteristic *zig–zag* pattern.

An atom in a given level can usually go one of two ways, and the proportion of times is goes one way or the other depends on the transition probabilities between the levels considered. I most complex atoms where the calculation of transition probabilities can be very difficult and we often have to resort to hard won experimental data which unfortunately often cannot cover all possible transitions. In the hydrogenic case the transition probabilities can be computed very precisely as follows:

The transition probability between two levels or states is understood in the wave function model as the extent of overlap of the probability wavefunctions for the two states. If two wavefunctions Ψ and Ψ' overlap, then the overlap integral or, matrix element, is

$$I = \int \Psi \mathbf{r} \Psi' d\mathbf{r}, \tag{2.13}$$

where $\mathbf{r}$ represents a spatial coordinate. This integral can be computed using a general purpose mathematical function called a hypergeometric function. The hypergeometric functions represent an infinite number of possible series, and change their behaviour dramaticly depending on what terms are used. In the case of the hydrogenic wavefunctions, the overlap integrals are particularly easy to handle using hypergeometric functions and infact turn out to reduce to ratios of simple polynomials. In a transition from an upper level n, l to a lower level n', l', where $l' = l \pm 1$, the overlap integral is given by,

$$I(n, l, n', l') = [c(n, n', l)F(n, n', l)]^2 \quad \text{if } l' = l - 1, \text{ or} \tag{2.14}$$
$$I(n, l, n', l') = [c(n', n, l')F(n', n, l')]^2 \quad \text{if } l' = l + 1. \tag{2.15}$$

Here

$$c(n, n', l) = \frac{(-1)^{(n'-l)}}{4(2l-1)!} \times \frac{((l+n'-1)!(l+n)!)^{1/2}}{((n'-l)!(n-l-1)!)^{1/2}}$$
$$\times \frac{(4nn')^{(1+l)}}{(n'+n)^{(n'+n)}} \times (n-n')^{(n+n'-2l-2)}, \tag{2.16}$$

and

$$F(n, n', l) = H[l - n + 1, l - n', 2l, (-4nn')/(n - n')^2]$$
$$- \frac{(n-n')^2}{(n+n')^2}H[l - n - 1, l - n', 2l, (-4nn')/(n - n')^2] \tag{2.17}$$

where H is the hypergeometric function $_2F_1(a, b, c, z)$. All this constitutes simple if messy algebra and H can be easily evaluated exactly using modern symbolic packages. Table 2.1.2 shows the overlap integrals for the transitions between the first 5 n levels of hydrogen. The integrals turn out to be exact rational numbers, and are given as the ratio of two, sometimes large, integers. To get an actual transition probability from the integrals the scalings that convert to physical units are required.

$$A(n, l, n', l', Z) = f[n, n', Z]^3 g[l, Z] I(n, l, n', l') \tag{2.18}$$

where Z is the atomic number and f and g are given by

$$f[n, n', Z] = cZ^2 R_{\mathrm{H}}(1/n'^2 - 1/n^2) \tag{2.19}$$
$$g[l, Z] = \frac{4l}{3(2l+1)} \left[\frac{h^3}{Z^2 c^3 m_e^2 \mathbf{e}^2} \right], \tag{2.20}$$

where c is the speed of light, R_H is the hydrogen Rydberg unit of energy, in modern units $R_H = 13.59844\,\text{eV}$ or $109678.774\,\text{cm}^{-1}$, h is Plank's constant, m_e is the electron mass and $\mathbf{e}$ is the electron charge in electrostatic units..

Table 2.1.2 also shows the evaluated transition probabilities for the low lying transition of hydrogen ($Z = 1$). While all these transitions are permitted or resonance lines, the transition probabilities do cover a wide range of values from 'typical' resonance values of $10^8 - 10^9\,\text{s}^{-1}$ down to values more typical of intercombination lines of only a few per second.

In order to compute the spectrum of hydrogen we need not only the transition probabilities, which serve to depopulate upper level and feed electrons down to lower levels, but we also need to have the recombination rates that feed electrons into this cascade of dropping electrons from outside the atoms. In addition the low density limit is sometimes does not apply and collisions between electrons and hydrogen atoms can move electrons back up the cascade and between l levels. Finally, when the radiation surrounding the hydrogen atom becomes strong enough at the right energies the photons can be absorbed and electrons can be again moved back up the cascade. It is the combination of collisions, recombination, absorption and transitions that makes the complete calculation of the hydrogen spectrum a challenge. We will look at these effects in subsequent chapters and build towards the means of computing the hydrogen spectrum and the emission from other species in diffuse plasmas. and transitions

2.1.3 The Spectroscopic Notation

The spectroscopic notation labels various transitions according to the state of the inner shell electrons, the electronic state of the outer optically–active electrons, and the quantum numbers of the resulting atomic state. However, students can find this initially very confusing. This is because the notation grew up, like many things in physics and astronomy initially as a set of observational phenomena, and only later was put on a sure theoretical basis. For example, in the nineteenth century, the observational effort was to discover regular series of lines which helped classify the plethora and confusion of emission lines seen through optical spectroscopes when individual elements were heated in flames or ionized in electrical discharges at low pressures. The brightest series were inevitably given the name "principal" (P), while the appearance of other sets of lines as seen through the spectroscopes was subtly different, and were accordingly given the names "sharp"(S), "diffuse"(D) or "fundamental"(F, on account of the hydrogen–like values of the series). It was only later realised that this classification had been according to the orbital angular momentum of the atomic configuration L; in order of increasing angular momentum from L = 0 these are: S, P, D, F. After that, imagination failed for the naming convention, and so the sequence continues: G, H, I....

As far as the wavefunctions of electrons are concerned, Schrödinger showed that, in the Coulomb potential, or indeed in other spherically symmet-

Table 2.1. Hydrogen transition probabilities between the first 5 levels.

Upper		Lower		Radial Integral	Probability	Observed	
n	l	n'	l'	$I(n,l,n',l')$	$A(n,l,n',l')\mathrm{s}^{-1}$	Line	$\lambda(\text{Å})$
2	1	1	0	$\left(\frac{32768}{19683}\right)$	6.258085E+08	Lyα	1215.67
3	1	1	0	$\left(\frac{2187}{8192}\right)$	1.670707E+08	Lyβ	1025.72
4	1	1	0	$\left(\frac{113246208}{1220703125}\right)$	6.811244E+07	Lyγ	972.54
5	1	1	0	$\left(\frac{625000}{14348907}\right)$	3.433791E+07	Lyδ	949.74
3	0	2	1	$\left(\frac{214990848}{244140625}\right)$	6.306708E+06	Hα	6562.80
3	1	2	0	$\left(\frac{2293235712}{244140625}\right)$	2.242385E+07	Hα	6562.80
3	2	2	1	$\left(\frac{27518828544}{1220703125}\right)$	6.458069E+07	Hα	6562.80
4	0	2	1	$\left(\frac{2097152}{14348907}\right)$	2.575344E+06	Hβ	4861.32
4	1	2	0	$\left(\frac{2621440}{1594323}\right)$	9.657538E+06	Hβ	4861.32
4	2	2	1	$\left(\frac{41943040}{14348907}\right)$	2.060275E+07	Hβ	4861.32
5	0	2	1	$\left(\frac{1728000000000}{33232930569601}\right)$	1.287221E+06	Hγ	4340.46
5	1	2	0	$\left(\frac{19906560000000}{33232930569601}\right)$	4.942930E+06	Hγ	4340.46
5	2	2	1	$\left(\frac{221184000000000}{232630513987207}\right)$	9.415105E+06	Hγ	4340.46
4	0	3	1	$\left(\frac{198429099687936}{33232930569601}\right)$	1.833427E+06	Paα	18751.0
4	1	3	0	$\left(\frac{994117681152000}{33232930569601}\right)$	3.061786E+06	Paα	18751.0
4	1	3	2	$\left(\frac{56358560858112}{33232930569601}\right)$	3.471578E+05	Paα	18751.0
4	2	3	1	$\left(\frac{1902101428961280}{33232930569601}\right)$	7.029945E+06	Paα	18751.0
4	3	3	2	$\left(\frac{243468982907044384}{232630513987207}\right)$	1.377295E+07	Paα	18751.0
5	0	3	1	$\left(\frac{516849609375}{549755813888}\right)$	9.037057E+05	Paβ	12818.1
5	1	3	0	$\left(\frac{701719453125}{137438953472}\right)$	1.635931E+06	Paβ	12818.1
5	1	3	2	$\left(\frac{64072265625}{274877906944}\right)$	1.493728E+05	Paβ	12818.1
5	2	3	1	$\left(\frac{2421931640625}{274877906944}\right)$	3.387776E+06	Paβ	12818.1
5	3	3	2	$\left(\frac{12109658203125}{1099511627776}\right)$	4.537200E+06	Paβ	12818.1
5	0	4	1	$\left(\frac{85762416640000000000}{4052555153018976267}\right)$	6.443709E+05	Brα	40512.0
5	1	4	0	$\left(\frac{2646238780456960000000}{36472996377170786403}\right)$	7.363837E+05	Brα	40512.0
5	1	4	2	$\left(\frac{338249646080000000000}{36472996377170786403}\right)$	1.882532E+05	Brα	40512.0
5	2	4	1	$\left(\frac{4938373529600000000000}{4052555153018976267}\right)$	1.484167E+06	Brα	40512.0
5	2	4	3	$\left(\frac{33554432000000000000}{1215766545905692801}\right)$	5.042185E+04	Brα	40512.0
5	3	4	2	$\left(\frac{2405181685760000000000}{1215766545905692801}\right)$	2.581599E+06	Brα	40512.0
5	4	4	3	$\left(\frac{343597383680000000000000}{109418989131512359209}\right)$	4.249545E+06	Brα	40512.0

Total Term Spin Multiplicity:
S is vector sum of electron spins (±1/2 each)
Inner full shells sum to 0

Term Parity:
o for odd, nothing for even

The Number of levels in a term is the smaller of (2S+1) or (2L+1)

Total Term Orbital Angular Momentum:
Vector sum of contributing electron orbitals.
Inner full shells sum to 0.

$2S+1 \, L_J^{(o)}$

$ns^i \, np^j \, nd^k...$

Electronic Configuration:
the electrons and their orbitals
(*i.e.* 1s2 2s2 3p1)

Total Level Angular Momentum:
Vector sum of *L* and *S* of a particular level in a term.

Fig. 2.3. Spectral notation for an atomic term, comprised of 1 or more levels.

ric potentials, it is possible to separate the eigenfunctions into their spatial parts. Thus in terms of the quantum numbers, the principal quantum number n, the angular momentum quantum number l, and the magnetic quantum number m, the wavefunction can be written in terms of the polar coordinates as:

$$\Psi_{nlm}(r, \theta, \phi) = R_{nl}(r)\Theta_{lm}(\theta)\Phi_m(\phi) \qquad (2.21)$$

where the quantum numbers satisfy:

$$\begin{aligned}
n &= 1, 2, 3, ..., \\
l &= 0, 1, 2, ..., (n-1) \text{ and} \\
m &= -1, -(l-1)..., 0, ...(l-1), l.
\end{aligned} \qquad (2.22)$$

The fact that the each electron has a spin angular momentum $s = 1/2$ ensures that the electron has a magnetic moment (equal to one Bohr magneton, $eh/4\pi m_e c$) which will interact with the magnetic field due to orbital motion, producing a total angular momentum $j = |l \pm 1/2|$.

Thus, for an orbital angular momentum $l = 0$, the only possible levels associated with the binding of a single electron have $j = 1/2; n = 1, 2, 3 ...$. In spectroscopic notation, these will be referred to as $n\ ^2S_{1/2}$ levels, meaning that they have principal quantum number n, the term is a doublet (although in this particular case the lower value of j would be negative, so one of the levels of the doublet cannot exist), the orbital angular momentum is zero (S state) and the total angular momentum is 1/2. Now consider the orbital

angular momentum $l = 1$. Now the possible levels have $j = 1/2, 3/2$ and $n = 1, 2, 3 \ldots$. In spectroscopic notation, these will be $n\ ^2P_{1/2}$ or $n\ ^2P_{3/2}$ levels. Continuing to higher orbital angular momentum, $l = 2$, $j = 3/2, 5/2$ and $n = 1, 2, 3 \ldots$ so these are $n\ ^2D_{3/2}$ or $n\ ^2D_{5/2}$ levels, and so on to higher l states.

Now, since the Pauli exclusion principle states that no two electrons can have identical quantum numbers, then the $n = 1$ state can be occupied by only two electrons ($l = 0, s = \pm 1/2$). This forms a closed shell of *configuration* $1s^2$. Here, the leading number is the shell number, (here the first), s refers to the angular momentum state of the electrons occupying this shell, and the superscript 2 refers to the number of electrons present. Thus the electron configuration of the normal state of magnesium ($Z = 12$) would be: $1s^2 2s^2 2p^6 3s^2$. Ionised magnesium, (Mg^+ or $Mg\,II$), which is isoelectronic with sodium and has one optically active electron in its outer shell will have the ground state, defined by the electron configuration and the ground term; $1s^2 2s^2 2p^6 3s\ ^2S_{1/2}$.

While one optically active electron only allows us one way of forming the total angular momentum by combining the spin and the orbital angular momentum, with two or more electrons life gets much more complicated. Consider the case of two electrons. For light atoms, as first shown by Russell & Saunders, the angular momentum vectors are coupled by electrostatic interaction to form the total orbital angular momentum, and the spins are also coupled by electrostatic interaction to give the total spin:

$$\mathbf{L} = \mathbf{l}_1 + \mathbf{l}_2 = (l_1 + l_2), (l_1 + l_2 - 1), \ldots |l_1 - l_2|$$
$$\mathbf{S} = \mathbf{s}_1 + \mathbf{s}_2 = 1/2 \pm 1/2 = 0, 1 \tag{2.23}$$

and the resultant $\mathbf{L}$ and $\mathbf{S}$ combine by magnetic interaction to give the total angular momentum:

$$\mathbf{J} = \mathbf{L} + \mathbf{S} = (L + S), (L + S - 1), \ldots |L - S| \tag{2.24}$$

This situation is known as (Russell–Saunders) **LS** coupling.

In the case of **LS** coupling for our atom with two optically active electrons, the possible states are $n\ ^1S_0$, $n\ ^1P_1$ and $n\ ^3P_{0,1,2}$. These are called the ground terms. In the case of neutral magnesium, $Mg\,I$, the ground state would be $1s^2 2s^2 2p^6 3s^2\ ^1S_0$, and the two excited states of the ground term are therefore $1s^2 2s^2 2p^6 3s^2\ ^1P_1$ and $1s^2 2s^2 2p^6 3s3p\ ^1P_{0,1,2}$. Clearly this is a rather clumsy notation, so usually the configuration of the closed shells is omitted, *viz.* $3s^2\ ^1S_0$, $3s^2\ ^1P_1$, and $3s3p\ ^1P_{0,1,2}$, respectively.

Other kinds of coupling are possible. For example, in heavier ions, the spin and orbit of individual electrons can be coupled, with the resulting total angular momenta of each electron being coupled together. This is called **JJ** coupling. This is merely indicative of the complexity that can be obtained with multi–electron atoms.

Table 2.2. C III Ion Transitions

Resonance Line: C III: A(2s 2p ^{1}P$_1$ - 2s^2 ^{1}S$_0$)	=	$1.7 \times 10^9 \, \text{s}^{-1}$
Intercombination Line: C III]: A(2s 2p ^{3}P$_1$ - 2s^2 ^{1}S$_0$)	=	$97 \, \text{s}^{-1}$
Forbidden Line [C III]: A(2s 2p ^{3}P$_2$ - 2s^2 ^{1}S$_0$)	=	$5.2 \times 10^{-3} \, \text{s}^{-1}$

2.1.4 Intercombination and Forbidden Lines

In pure **LS** coupling, the wavefunctions for the two states which obey the selection rules given in section (2.1.1) are strictly orthogonal, so that the overlap integral, the line transition probability and the gf value are all strictly zero. However, in real ions of intermediate atomic mass, the coupling is not pure **LS**, but is usually inermediate between **LS** and **JJ**. For heavy atoms, other coupling schemes may be a closer approximation to reality. The departure from pure **LS** coupling means that electric quadrupole transitions between states of different multiplicity can occur, albeit with low transition probability. These are the so–called *intercombination, inter–system or semi–forbidden lines.* To distinguish these transitions from normal resonance transitions, we usually place a square bracket to the right of the ion identification, *i.e.* C III]. Typically, the transition probabilities are a factor of order $\alpha^2 = \left(\mathbf{e}^2/hc\right)^2$ lower than a resonance line of the same frequency. Since resonance lines have radiative transition probabilities of order $10^8 - 10^9 \, \text{s}^{-1}$, then intercombination lines have $A \sim 10^3 \, \text{s}^{-1}$. For the typical densities and temperatures found in the interstellar medium, the intercombination line transition probability is still comfortably larger than the mean atom–atom collisional rate.

In nebulae at low enough densities or temperatures, we can even see magnetic quadrupole transitions. These are even more highly improbable, with transition probabilities of order $\alpha^4 = (\mathbf{e}^2/hc)^4$ lower than a resonance line at the same frequency. These are called *forbidden lines*. Examples of all these transitions are to be found in the C III ion in table 2.1.4:

Another example is illustrated by the partial Grotian (level) diagram for the O III ion shown in Figure (2.4), where the wavelengths are given in angströms, (Å). Although electrons involved in forbidden transitions can rest in their excited states for several minutes or more, the very low densities of diffuse astrophysical plasmas ensure that even over such long periods of time, a collision between the ion and an electron may be quite unlikely, so that the atom is left to radiate the forbidden line. The photon produced has a negligible probability of being re–absorbed by another ion thanks to the low transition probability (*c.f.* equation (**??**)), so it is free to escape the region where it is being excited. In this way, despite their improbability, forbidden lines are very important in setting the energy balance of diffuse astrophysical plasmas.
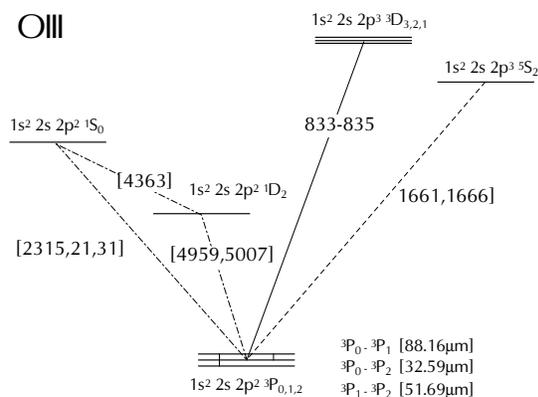
Fig. 2.4. A partial Grotian (energy level) diagram for the lower states of the O III ion, illustrating several classes of transition which are observed in diffuse astrophysical plasmas.

2.2 Molecular Spectra

In the last twenty years there has a been a remarkable development in the field of interstellar chemistry. This has come about as the result of detailed laboratory measurements of the rotation–vibration emission spectra of candidate molecules in the microwave region of the spectrum, and the astronomical study of microwave lines from dense and warm clouds of molecular gas. The famous source in the Galactic Centre, Sgr B2 has been a favourite hunting ground for new molecules, but many molecular species have been detected in other regions around newly–formed hot stars, such as the Orion Nebula Kleinemann–Low region, and the compact H II regions W51 and DR21. Molecules as complex as carbon chains with 13 atoms have been detected, and the number of species detected is increasing weekly. A list of species that have been detected by mid-1998 is given in Table 2.2:

The observational material is now sufficient that many of the details of the astrochemical evolution of the interstellar medium are now being elucidated. For those that want to know more the monograph by Emma Bakes (1997) gives an excellent introduction to the field.

The field of molecular spectroscopy is very complex, and we cannot hope to do justice to it here, but we hope simply to summarise some of the nomenclature that is used so that the field can be more easily approached.

Table 2.3. Molcular Species Detected in the ISM (after Snyder(1997) and Bakes(1997))

Atoms	Species
2	H_2 HD C_2 N_2 OH CH CN CO CP CS NH NO NS SO PN AlF AlCl HCl KCl NaCl SiC SiN SiO SiS CH^+ CO^+ CN^+ NO^+ MgH^+ SO^+ CS^+
3	C_3 C_2H CH_2 C_2O C_2S CO_2 HCN HCO NH_2 N_2O H_2O H_2S HNC HNO MgCN MgNC NaCN OCS SO_2 c-SiC_2 HCO^+ HCS^+ H_3^+ HOC^+ N_2H^+
4	c-C_3H i-C_3H C_3N C_3O C_3S C_2H_2 HCCN HNCO HNCS H_2CO H_2CN CH_2N H_2CS NH_3 CH_2D^+ $HCNH^+$ $HOCO^+$ H_3O^+
5	C_5 C_4H C_4D l-C_3H_2 c-C_3H_2 CH_2CN CH_4 HC_3N HC_2NC CH_2C_2 HCOOH H_2CNH H_2C_2O H_2NCN HNC_3 SiH_4 C_4 Si H_2COH^+ HC_3NH^+
6	C_5H l-H_2C_4 C_2H_4 C_5S? CH_3CN CH_3NC CH_3OH CH_3SH HC_2CHO $HCOC_2H$ $HCONH_2$ HC_3NH^+
7	C_6H HC_5N CH_2CHCN CH_3C_2H HC_5N $HCOCH_3$ NH_2CH_3 c-C_2H_4O CH_3CHO
8	C_7H H_2C_6 CH_3COOH CH_3C_3N $HCOOCH_3$
9	C_8H HC_7N CH_3C_4H CH_3CH_2CN $(CH_3)_2O$ CH_3CH_2OH
10	CH_3C_5N? $(CH_3)_2CO$
11	HC_9N
12	
13	$HC_{11}N$

2.2.1 Rotating Diatomic Molecules

The simplest molecular line spectra are due to diatomic molecules in rotational motion. These molecules emit in the infrared and microwave regions of the electromagnetic spectrum and produce a spectrum of emission or absorption lines which are nearly equally spaced in frequency (or energy). These lines are formed by transitions between quantized rotational energy levels which are directly related to the masses and inter–atomic spacing of the atoms in the molecules.

To a first approximation a slowly spinning diatomic molecule can be considered as a rigid rotator – that is, the interatomic spacing is constant. In reality, as the molecules spin faster centrifugal effects can change the atomic spacing and hence the energy level spacing. Fortunately, many molecular

species such as H_2 detected in diffuse astrophysical plasmas are observed in the lowest spin levels where the rigid rotator model is a fair approximation to the truth.

The rotation of a rigid diatomic rotator is equivalent to the motion of a single reduced mass μ on a sphere of radius r from the centre of rotation, analagous with the hydrogen atom discussed above. The rotator is a bit simpler though, because while the reduced system is spherical, it has a fixed radius. The wavefunctions can be separated into $\Theta(\theta)$ and $\Phi(\phi)$ angular components in spherical coordinates as usual, but the radial component disappears and need not be included.

The solutions are quantised in two variables, m_j the quantum number of the z component of the angular momentum and J the rotational quantum number (analogous to l in one electron atoms). For the m_j the simplest determining expression is:

$$\frac{d^2\Phi}{d\phi^2} = -m_j^2\Phi, \tag{2.25}$$

The quantum number J is related to the moment of inertia, μ and r^2, and the rotational energy E by:

$$\frac{8\pi^2\mu r^2 E}{h^2} = J(J+1) \tag{2.26}$$

and the wave equation constrains J to integral values and $J \geq |m_j|$, so there are $2J + 1$ values of m_j for each value of J.

Some of the rigid rotator wavefunctions are represented in figure (2.3) for $J = 0$ up to $J = 4$. The complex wavefunction Ψ is represented as two rectangular ϕ - θ maps, analogous to a mercator projection world map. This is done in preference to the more usual polar line plots because it shows the symmetry properties of the wavefunctions more clearly. The probability density $\Psi^*\Psi$ is shown, and black represents the regions where the rotating particle spends the most time as it spins. The high J, high $|m_j|$ plots show the trend towards the classical high angular momentum result with the nuclei spinning in a well defined equatorial plane. The probability plots are also mapped onto a spherical representation to aid visualising the molecule's motion.

From equation (2.26), the possible rotational energy levels in the rigid rotator are:

$$E_{\text{rot}} = J(J+1)\frac{h^2}{8\pi^2 I}. \tag{2.27}$$

Transitions between different J levels are possible through absorption and emission of photons. However, since photons have a spin angular momentum of $h/2\pi$, and so the conservation of angular momentum limits the change in angular momentum of the molecule to $\Delta J = \pm 1$. In collisional processes, the colliding particle can inject or carry away any amount of angular momentum, so here transitions with $|\Delta J| > 1$ are possible. The energy of a transition is given by,

$$E' - E'' = \Delta E = \frac{h^2}{8\pi^2 I}[J'(J'+1) - J''(J''+1)]. \qquad (2.28)$$

When $J' = J'' + 1 = J + 1$, for radiative transitions, this reduces to

$$\Delta E = \frac{h^2}{4\pi^2 I}(J+1), \qquad (2.29)$$

and the lines frequencies are,

$$\nu = \frac{\Delta E}{h} = \frac{h}{4\pi^2 I}(J+1) = 2B(J+1), J = 0, 1, 2... \qquad (2.30)$$

where B is the commonly used molecular rotation constant $h/(8\pi^2 I)$. Equation (2.30) implies that the successive J transitions are separated by a constant value in frequency; $\Delta\nu = 2B$. Astrophysical examples of such rotational ladders are presented by the CO or CS molecules. In CO the observed wavelengths of the $J = 1 - 0$, $J = 2 - 1$ and $J = 3 - 2$ transitions are 115271.204 MHz, 230537.974 MHz and 345795.900 MHz, respectively. If the separations were rigorously identical, then we would expect the $J = 2 - 1$ and $J = 3 - 2$ transitions to occur at 115271.204 MHz, 230542.408 MHz and 345813.612 MHz, respectively. The slight lowering of the observed frequency compared with that predicted is due to the centrifugal distortion mentioned above. This increases the separation of the molecules, and hence the moment of inertia, and so lowers the energy of the transition, as observed. The change in the moment of inertia of the molecule obtained by substituting a ^{13}C for the ^{12}C in the CO molecule shifts the frequency of the $J = 1 - 0$ transition to 110201.370 MHz (c.f. 110204.338 MHz expected simply from the ratio of the reduced masses of the two molecules).

2.2.2 Rotational Wavefunction Symmetry

Advanced Section

In addition to angular momentum conservation, considerations of symmetry places important restrictions on the allowable rotational transitions. In any quantum wavefunction system, such as an atomic or molecular system, the wave equation (Eqn **??**) should remain unchanged after reflection about the origin (x, y, z replaced by $-x$, $-y$, $-z$), that is, it should not depend on the choice of coordinate systems. However the eigenfunctions, or solutions, as shown in Figure (2.5) may be affected by reflection about the origin, and will be either unchanged or have a sign reversal after the reflection. To see this, take a wavefunction in Figure (2.4), $\Psi(real)$ for $J = 3$ and $|m_j| = 2$ for example, flip it vertically and then shift it left to right by half the width of the diagram with wrap around. Doing this results in the black and white pattern exactly inverting. This process is the graphical equivalent of reflection about the origin point, that is $\theta' = \pi - \theta$ and $\phi' = \pi + \phi$. Inspection will show that all the wavefunctions for even values of J are identical (symmetric) after this
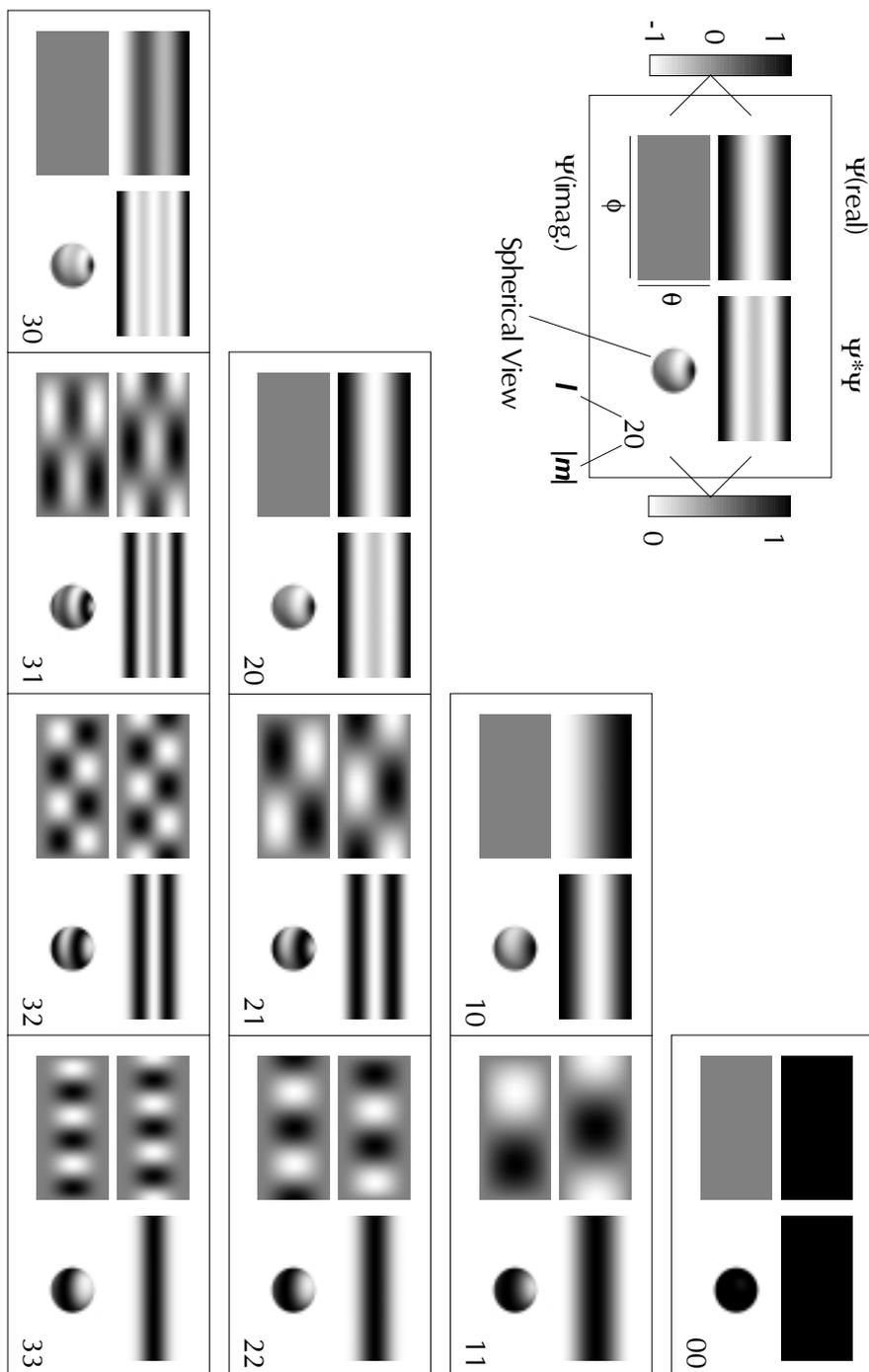
Fig. 2.5. Angular spherical wavefunctions.

process, and all the wavefunctions for odd values of J invert or change sign, and are thus anti–symmetric.

The rotational eigenvalues or levels are classified as either *positive* or *negative* depending on the symmetry of the *total* wavefunction under reflection,

$$\Psi_{\text{tot.}} = \Psi_e \left(\frac{1}{r}\right) \Psi_\nu \Psi_{\text{rot.}} \tag{2.31}$$

this total wavefunction being a combination of nearly independent electronic (Ψ_e), vibrational ($(1/r)\Psi_\nu$) and rotational ($\Psi_{\text{rot.}}$) wavefunctions. In *positive* rotational levels $\Psi_{\text{tot.}}$ remains unchanged on reflection and for *negative* rotational levels $\Psi_{\text{tot.}}$ change sign. This is analogous to even and odd atomic energy levels.

Now, the vibration wavefunction is only a function of separation r, and is unchanged by reflection. In diatomic molecules, the net angular momentum about the inter–nuclear axis is typically zero ($\Lambda = 0$), although there are exceptions, so the electronic wavefunction (Ψ_e) is also typically symmetric under reflection. In this case the *parity*, *positive* or *negative*, will be determined by the symmetry of the rotational wavefunctions alone. Positive levels are symmetric and negative levels are anti–symmetric. If Ψ_e is anti–symmetric then the reverse is true and positive levels are anti–symmetric rotation levels and negative levels are symmetric. This principle is true even if the total wavefunction is not perfectly separable as implied in eqn (2.31).

Using symmetry, a selection rule for transitions between positive and negative rotational levels can be found by considering the symmetry of the quantum mechanical operator associated with the transition. In the case of electric dipole transitions, the dipole operator $\mathbf{D}$ is a vector sum of charge displacements, and point reflection will change the vector direction to exactly the opposite, and $\mathbf{D}$ is therefore anti–symmetric under this reflection. The transition probabilities between two states Ψ_1 and Ψ_2 are determined by the associated operators (in general $\mathbf{O}$) for the transition in so–called matrix elements, which are integrals over all space $d\tau$,

$$P(1-2) = \int \Psi_1 \mathbf{O} \Psi_2 d\tau. \tag{2.32}$$

The integral, under situations where symmetry is present, can be one of two alternatives:

1. When the integrand $\Psi_1 \mathbf{O} \Psi_2$ is symmetric the integral can take on values other than zero.
2. When the integrand $\Psi_1 \mathbf{O} \Psi_2$ is anti–symmetric the integral will be exactly zero.

The symmetry of the integrand is the product of the symmetry of the terms. If the operator $\mathbf{O}$ is anti–symmetric like the dipole operator $\mathbf{D}$ then the symmetry of the integrand is determined using simple product rules. Let

S be any symmetric function and A be any anti–symmetric one, then the following apply:

$$SS \to S$$
$$SA \to A$$
$$AS \to A$$
$$AA \to S$$

So, for the transition probability to be non–zero, with an anti–symmetric operator like $\mathbf{D}$, the wavefunctions Ψ_1 and Ψ_2 must be of opposite symmetry. This can be confirmed by taking triplets of S s and As with the middle term being A and multiplying through with the above rules.

In the rotator case, all the *positive* levels are either symmetric or anti–symmetric and similarly for the *negative* levels. So the only transitions that are allowed under the dipole operator are *positive* to *negative* and *negative* to *positive* transitions. This is more general but is still consistent with the $\Delta J = \pm 1$ rule derived earlier. Other operators such as for magnetic dipole or quadrupole transitions can be symmetric under point reflection and thus have the opposite selection rule, however they are still restricted by angular momentum restrictions.

2.2.3 Rotating Diatomic Molecules with Identical nuclei

If the two nuclei of the diatomic molecule are identical, such as in the astrophysically important H_2 molecule, further more rigorous symmetry occurs. In addition to reflection symmetry about the origin point, there is symmetry under the exchange of the two nuclei – or reflection about a plane between the nuclei. The total wavefunction must remain unchanged or change sign after the exchange of the two nuclei. The states are either *symmetric* or *anti–symmetric in the nuclei* respectively.

For a given electronic state, Ψ_e, either all the positive (in origin reflection) rotational states are symmetric (in the nuclei) and the *negative* levels are anti–symmetric or the opposite is true. If Ψ_e does not change sign on exchange, and $\Lambda = 0$, then the states with *positive symmetric* and *negative anti–symmetric* levels are designated Σ_g^+ and those with *negative symmetric* and *positive anti–symmetric* levels are designated Σ_u^+. If the Ψ_e does change sign on exchange, then the equivalent states are Σ_u^- and Σ_g^-.

If the nuclei have zero spin, or nuclear spin is ignored, then there is a very strict selection rule in addition to the previous ones. Here, transitions between symmetric and anti–symmetric states are completely forbidden, for not only any radiative transitions, but also for all collisional and other processes as well.

First consider dipole radiative transitions. Previously we had the dipole operator $\mathbf{D}$ in the matrix element integrals anti–symmetric under point origin

reflection, forcing the product of the two level wavefunctions Ψ_1 and Ψ_2 to be anti–symmetric under reflection so the integral would be symmetric and hence able to have a non–zero value. This could only be achieved if the levels were positive and negative or emphvise versa. Under nuclear exchange, or plane reflection symmetry, the dipole operator $\mathbf{D}$ is always symmetric, as is any other multipole. Therefore, for the integrand $\Psi_1 D \Psi_2$ to be symmetric, and the integral be able to have values other than 0, Ψ_1 and Ψ_2 must have the same symmetry. If they do not then the integral will be 0 and the transition forbidden. Even in collisional processes, any third particle interacting with the diatomic nucleus cannot distinguish between the exchange of the nuclei, so the collisional processes are *symmetric in the nuclei*, and so cannot perform transitions between symmetric and anti–symmetric states. Since symmetry alternates between symmetric and anti–symmetric levels in J, ΔJ is *never* an odd number of levels.

The implication of this is that systems of zero nuclear spin, like O_2, and C_2, could exist in two distinct populations, one with only symmetric rotation levels, and one with anti–symmetric rotation levels, depending on what symmetry the molecules were formed with. These molecules would not be seen directly in dipole transitions between rotational levels, but may be seen in rotational fine structure of the vibration lines of electronic transitions. Observationally *only* the odd J levels are occupied, which are symmetric, and no anti–symmetric rotational levels are present. In Raman spectra of O_2, where transitions occur with $\Delta J = 0$ or ± 2 and thus are allowed, all the even numbered lines are missing, supporting the hypothesis that while anti–symmetric levels are mathematically possible, in our universe only the symmetric ones are used. This is presumably related to the intrinsic symmetry of the nuclei before the molecule formation and the symmetry of the process that forms them.

In the case of H_2, the nuclei have non–zero spin – namely $\pm 1/2$. The nuclear spin of the molecule can take on values according to the vector sums of the two spins, either 1 (both $+1/2$, or *parallel, anti–symmetric*) or 0 (one $+1/2$ and the other $-1/2$, *anti–parallel , symmetric*). More generally for various spins, the total nuclear spin can take on the values,

$$N = 2I, 2I - 1, ..., 0 \qquad (2.33)$$

where I is the spin number of each nucleus. Also each value of N has a statistical weight of $2N + 1$ in the usual manner. For H_2, $I = 1/2$ and the *parallel* state has a statistical weight of 3 and the *anti–parallel* a weight of 1. Furthermore, when the nuclear wavefunction is combined with the rest of the molecule wavefunction in eqn (2.31), the symmetry of the nuclear wavefunction can change the symmetry of the overall wavefunction. For example, an anti–symmetric nuclear function multiplied by a symmetric molecular function will become anti–symmetric overall. However this alone is not sufficient to allow symmetric to anti–symmetric transitions to occur. If the nuclear wavefunction Ω and molecular wavefunction Ψ are separable:

$$\Psi_{\text{tot.}} = \Psi \Omega \tag{2.34}$$

In this case the transition matrix element integrals are also separable, and each part must still obey the symmetry laws and be strictly zero for symmetric to anti–symmetric states. However if the nuclear function is coupled to the molecular wavefunction, so that some part ϵ is not separable,

$$\Psi_{\text{tot.}} = \Psi \Omega + \epsilon. \tag{2.35}$$

Then the symmetry of the rotational components of the total wavefunction are not perfectly symmetrical and the transition probabilities will be non–zero. The nuclear coupling is usually extremely weak however, and the transition rates are very small indeed. The mean lifetime of a molecule of one rotational symmetry before it makes a transition to a state of another symmetry ($\Delta J = \pm 1$) can be years.

As a result of this there are two types of H_2 molecules; one with anti–symmetric (parallel) nuclear and anti–symmetric rotational wavefunctions (odd J numbers), and one with symmetric (anti–parallel) nuclear and symmetric rotational wavefunctions (giving even J numbers). Because of their long lifetimes, these behave as two separate rotation systems which can only very slowly mix with one another. Given enough time - and that is a lot, they will come into statistical equilibrium with each other in the ratio of their statistical weights, that is $3 : 1$. The most common type is known as the *ortho-* variety, ortho–hydrogen or $o-H_2$, and the less common variety is *para*-hydrogen or $p-H_2$. In homonuclear molecules with integral spins, *i.e.* deuterium, D_2, with $I = 1$, a similar circumstance arises, with the most common variety also known as ortho–deuterium, however in this case the ortho variety are the symmetric systems and the para variety is the less common anti–symmetric type.

Notes on Chapter 2

- A very useful introductory summary of quantum physics is given by James William Rohlf *"Modern Physics from α to Z^0"*, John Wiley & Sons: NY, 1994.
- The quantum mechanics of molecules is fully treated in the book by P.W. Atkins *"Molecular Quantum Mechanics"* Second Edition, 1983, (Oxford University Press: Oxford), ISBN 0-19-855170-3.
- The atomic and molecular physics relevant to IR and radio astronomy is clearly summarised by Reinhard Genzel in Saas–Fee Advanced Course 21, *"The Galactic Interstellar Medium"*, 1991, (Springer–Verlag: Berlin), ISBN 3-540-55805-5, whose work has been used in the preparation of this chapter and in Chapter 4.

- The key reference work of atomic terms, configurations and energy levels is *"Grotrian Diagrams"*.

Exercise 2.2.1. What makes the following transition forbidden: $^1D_2 \rightarrow$ $^3P_2$?

Exercise 2.2.2. How many optically–active electrons does the O^{2+} (or O III) ion contain? What is the electron configuration of the ground terms? What are the multiplets of the ground terms?

Exercise 2.2.3. The $J = 1 - 0$ transition in ^{12}CO occurs at 115271.204 MHz. Calculate the interatomic separation of this molecule in Å (10^{-8} cm). Assume the mass of a hydrogen atom is 1.66×10^{-24} g.

3. Collisional Excitation

"High o'er the fence leaps Sunny Jim,
'Force' is the food that raises him"
— Minnie Hanff (Advertising slogan, 1903)

Under the density conditions that are found in most of the interstellar medium, most of the atoms of any particular element and ionization state reside in the ground state. The excited states of atoms are mostly populated as a result of collisions with charged species (usually electrons or protons), and the *collision excitation* timescale is often long enough that the atoms are free to decay back to the ground state by radiative processes. A low enough densities, this condition will remain true even if the excited state has a radiative lifetime of several seconds. This is freqently the case for the forbidden transitions observed in ionized plasmas.

In these conditions, the flux produced in an emission line resulting from the radiative decay of the excited level will depend on the number of collisions, which is proportional to the product of the number densities of the two colliding species multiplied by the probability that a collision will produce a collisional excitation. In turn, this probability depends on the collisional excitation cross section integrated over the energy distribution of the colliding species; determined by the temperature T. It is almost intuitive that if the energy gap between the ground state and the excited state, E_{12}, is much larger than the mean energy of the colliding species $\sim kT$, then, because there are few very energetic collisions, few collisional excitations can occur and the resulting line will be very much weaker than when $E_{12} \leq kT$. This gives us the possibility of measuring temperature from the relative strengths of lines coming from different excited levels. The same kind of thing can happen in molecules. For example, molecular collisions can excite various rotational states of molecular hydrogen, so a comparison of the column densities in the various rotationally-excited levels can be used to estimate the kinetic temperature.

At high enough densities, the collision timescales are short, and the population in any upper level is set by the balance between collision excitation, and the *collision de-excitation* out of these levels. Under these conditions, the populations in the excited states are governed by the Boltzmann equilibrium; equation (1.1). At intermediate densities where the collisional rates and the
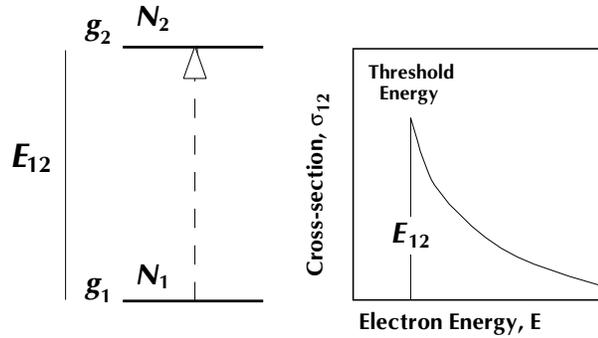
Fig. 3.1. Above the threshold energy for collisional excitation, the cross section decreases approximately inversely as energy.

radiative decay rates are comparable, the intensity of an emission line is determined both by the temperature and the density. Thus if the temperature is known, the density can be determined from the intensity ratio of two such lines. These simple considerations hold the key to the spectroscopic determination of the physical conditions in any given parcel of interstellar plasma. Let us now consider the physics of collisional excitation in more detail.

3.1 Collisional Excitation by Electron Impact

Consider an (idealised) atom with only two energy states, a ground state and an excited state which can radiate back down to the ground state. Electron impacts can collisionally excite an atom into an upper level. Once there, if left alone, it will return to the ground state by a radiative transition. If, on the other hand, it suffers another collision with an electron while still in the excited state, it may collisionally de–excite back down to the ground state. The collisional cross–section is a strongly varying function of energy, in general varying approximately inversely as the impact energy.

Therefore, the collision strength, Ω_{12}, defined in terms of the collisional cross–section, $\sigma_{12}(E)$:

$$\sigma_{12}(E) = \left(\frac{h^2}{8\pi m_e E} \right) \left(\frac{\Omega_{12}}{g_1} \right) \text{ cm}^2, \tag{3.1}$$

is a more convenient quantity to use, since it removes the primary energy dependence for most atomic transitions. In this equation, m_e is the electron mass, E is the electron energy, and g_1 is the statistical weight of the ground state. In addition Ω_{12} has another advantage; that of symmetry between the upper and the lower states. This is readily demonstrated. Consider the situation at very high density, where the population of the ground and the excited level are determined entirely by collisions, the radiative rate being negligible by comparison to either the collision excitation rate or the collisional de–excitation rate. These circumstances ensure that the atom is in Local Thermodynamic Equilibrium (LTE) and that the two levels are populated according to the Boltzmann equilibrium (equation 1.1) at the electron temperature, T, (equation 1.2):

$$\frac{N_2}{N_1} = \frac{g_2}{g_1} \exp\left[\frac{-E_{12}}{kT}\right] \tag{3.2}$$

Consider the equation of detailed balance at high density (Boltzmann Equilibrium). However we can also compute the population ratio from equation (3.1), using the *principle of detailed balance.* This states that, in equilibrium, the rate of population of the upper level per unit volume, through collisional excitation, R_{12} ($cm^{-3}\,s^{-1}$) is equal to the rate of depopulation through collisional de–excitation, R_{21} ($cm^{-3}\,s^{-1}$). If the electron density is n_e (cm^{-3}), and these are distributed according to the Maxwell–Boltzmann energy distribution function $f(E)dE$; given by equation (1.2), then:

$$R_{12} = n_e N_1 \int_{E_{12}}^{\infty} \sigma_{12}(E).E.f(E)dE$$
$$= n_e N_1 \, \alpha_{12} \tag{3.3}$$
$$= n_e N_1 \left(\frac{2\pi\hbar^4}{km_e^3}\right)^{1/2} T^{-1/2} \left(\frac{\Omega_{12}}{g_1}\right) \exp\left[\frac{-E_{12}}{kT}\right] \; cm^{-3}\,s^{-1},$$

and

$$R_{21} = n_e N_2 \int_{0}^{\infty} \sigma_{21}(E).E.f(E)dE$$
$$= n_e N_2 \, \alpha_{21} \tag{3.4}$$
$$= n_e N_2 \left(\frac{2\pi\hbar^4}{km_e^3}\right)^{1/2} T^{-1/2} \left(\frac{\Omega_{21}}{g_2}\right) \; cm^{-3}\,s^{-1}.$$

(it is a useful exercise to derive these two equations). The alphas, α_{12} amd α_{21} are known as the collisional excitation and de–excitation coefficients, have units of $cm^3\,s^{-1}$, and in general $\alpha_{12} \neq \alpha_{21}$ because of the Boltzmann factor $\exp[-E_{12}/kT]$ and different statistical weights.

Applying detailed balance by setting the rate in equation (3.3) to that of equation (3.4) gives the ratio of the two level populations in terms of the rate corefficients:

$$\frac{N_2}{N_1} = \frac{\alpha_{12}}{\alpha_{21}} = \left(\frac{\Omega_{12}}{g_1}\right)\left(\frac{g_2}{\Omega_{21}}\right)\exp\left[\frac{-E_{12}}{kT}\right] \tag{3.5}$$

Comparing this with the Boltzmann Equation (eqn 3.2), it follows that the collision strength, Ω, has to be symmetric between the levels, *i.e.* $\Omega_{12} = \Omega_{21}$.

There is a simple quantum mechanical sum rule for collision strengths for the case that one term consists of a single level, and the second consists of a multiplet. This occurs, for example, in the case that one of the levels has had its degeneracy removed by spin–orbital interactions, as in, for example, the individual levels within the term $^3P_{0,1,2}$. For such terms:

$$\Omega_{(L_1,S_1.J_1:L_2,S_2.J_2)} = \frac{(2J_2 + 1)}{(2S_2 + 1)(2L_2 + 1)}\Omega_{(L_1,S_1:L_2,S_2)} \tag{3.6}$$

provided that either $S = 0$ or $L = 0$. Here, the factor $(2J_2+1)$ is the statistical weight of an individual level in the multiplet, and $(2S_2 + 1)(2L_2 + 1)$ is the statistical weight of the multiplet. In the example of the term $^3P_{0,1,2}$, we can regard the collision strength as being "shared out" amongst these levels in proportion to the statistical weights of the individual levels; $g_J = (2J + 1)$. Thus, the $^3P_0$ level will carry 1/9 of the total collision strength, the $^3P_1$ level has a fraction 3/9, and the $^3P_2$ level accounts for 5/9 of the total.

Quantum mechanical calculations show that the resonance structure in the collision strengths is important, and that, for neutral species, the collision strength increases with energy. However, because of the property of resonances that they oscillate -ve and +ve over a small energy range, then the effect on the temperature averaged collision strength (average over the Maxwell–Boltzmann distribution), Ω_{ij} is usually small, and can be usually fitted by a simple power law; $\Omega_{ij} = A + B(T/10^4K)^n$. For electric dipole transitions, the collision strength and the gf value of the transition are related through the equation given by Seaton (1958):

$$\Omega_{ij} = \left(\frac{8\pi}{\sqrt{3}}\right)E_{ij}^{-1}gf_{ij}G(T) \tag{3.7}$$

where $G(T)$ is a *Gaunt Factor*, which is a numerical multiplication factor which changes the result of a "classical" calculation into one which is rigorously identical to the result produced by a full quantum mechanical calculation (often a 'fudge' factor or an empirical value when the full quantum solution is not solved yet). In this case, the $G(T)$ is a fairly complex function of temperature. This can be written (from Landini & Monsignori Fosse, 1991) in terms of the first exponential integral E_1 as:

$$G(x) = A + \exp[x]\left(Bx - Cx^2 + Dx^3 + E\right)E_1(x) + (C + D)\,x - Dx^2. \tag{3.8}$$
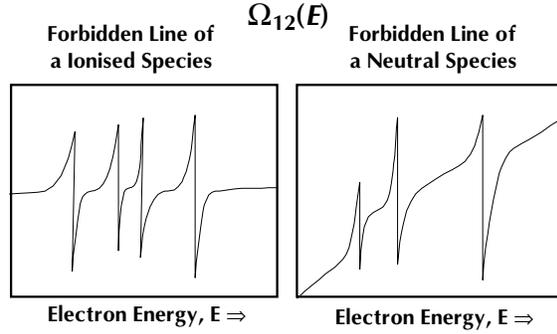
Fig. 3.2. Computed variation of the collision strength for an ionised and a neutral species, taken from the work of the Opacity Project (REF**)

However, for simple approximate purposes, a temperature averaged Gaunt factor can be substituted in equation.(3.7) to give

$$\Omega_{ij} = \frac{8\pi}{\sqrt{3}} \left(\frac{I_H}{E_{ij}} \right) g f_{ij}, \tag{3.9}$$

where here I_H is the ionisation potential of hydrogen.

3.1.1 Limiting Cases

Let us now consider some important limiting cases of the two level atom. First, in the low density limit, the collisional rate between atoms and electrons is much slower than the radiative de–excitation rate of the excited level. Thus we can balance the collisional feeding into level 2, given by equation (3.3) by the rate of radiative transitions back down to level 1; $R_{12} = A_{21}N_2$, where A_{21} is the radiative transition probability downwards. Balancing these two rates gives:

$$N_2 = n_e N_1 \frac{\alpha_{12}}{A_{21}} \tag{3.10}$$

$$= n_e N_1 \, \beta \, A_{21}^{-1} T^{-1/2} \left(\frac{\Omega_{12}}{g_1} \right) \exp \left[\frac{-E_{12}}{kT} \right] \ \ \text{cm}^{-3}$$

where β is the collection of constants: $\beta = \left[(2\pi\hbar^4)/(km_e^3)\right]^{1/2}$, which has the value 8.62942×10^{-6} in c.g.s units. The emission line flux can then be computed, since it is the product of the number of transitions occurring per unit time and the energy of each photon, $F_{12} = E_{12}A_{21}N_2$. Substituting equation (3.10) into this, we have;

$$F_{12} = \chi_i n_e^2 \beta E_{12} T^{-1/2} \left(\frac{\Omega_{12}}{g_1}\right) \exp\left[\frac{-E_{12}}{kT}\right] \; \mathrm{erg\, cm^{-3}\, s^{-1}} \qquad (3.11)$$

where χ_i is the relative abundance (by number) of the ion considered; $N_i = \chi_i n_e$. The form of this is shown in fig **??** (below). Note that, for low temperatures, the exponential term dominates because few electrons have energy above the threshold for collisional excitation, so that the line rapidly fades with decreasing temperature. At high temperatures, the $T^{-1/2}$ term controls the cooling rate, so the line fades slowly with increasing temperature. From equation (3.11), it is evident that the line flux reaches a maximum at a temperature $T = E_{12}/k$.

In the high density limit, the level populations are set by the Boltzmann equilibrium equation (3.2), so that the line flux is now:

$$F_{12} = \chi_i n_e E_{12} A_{21} \left(\frac{g_2}{g_1}\right) \exp\left[\frac{-E_{12}}{kT}\right] \; \mathrm{erg\, cm^{-3}\, s^{-1}}. \qquad (3.12)$$

Here, the line flux scales as n_e rather than n_e^2, but the line flux tends to a constant value at high temperature. The change of the power of density dependence implies the existence of a *critical density*, $n_{\mathrm{crit.}}$, defined as the density where the radiative depopulation rate matches the collisional de–excitation for the excited state; $A_{21} = R_{21}$, *i.e.*

$$A_{21}N_2 = \beta(\Omega_{12}/g_2)T^{-1/2}n_{\mathrm{crit.}}N_2. \qquad (3.13)$$

from which:

$$n_{\mathrm{crit.}} = \left(\frac{A_{21}g_2 T^{1/2}}{\beta\Omega_{12}}\right) \mathrm{cm}^{-3} \qquad (3.14)$$

At around this density, the line emissivity plotted on log : log coordinates changes slope from $+2$ to $+1$, and so this density can be considered to represent the transition from the low density limit to LTE populations for this particular transition.

3.2 The Three–Level Atom

Although all ions have many excited states, nonetheless nature provides a large number of atoms which, for ISM studies can be effectively treated as three–level systems. Generally, these involve transitions between different terms (including the ground term) having the same principal quantum

number. As a consequence, many of the transitions that are involved are
forbidden, and are only important at the low densities characteristic of the
ISM. However, many of these transitions occur at optical wavelengths, and
counted amongst the strongest lines in ground–based spectra of ionised nebu-
lae. Such systems provide an instructive insight into the main way line ratios
can be used for plasma diagnostics for the ISM, and have given us our basic
understanding of the physical conditions applying in ionised regions in the
ISM.

Let C_{ij} be the collision rate $(C_{ij} = n_e \alpha_{ij} \, \mathrm{s}^{-1})$ between any two levels
defined by equations (3.3) and (3.4), above; depending on whether the rate
represents a collisional excitation or a collisional de–excitation. If the A_{ij} are
the radiative transitional rates, then the equations of statistical equilibrium
for a three–level atom are:

$$
\begin{aligned}
N_1 C_{13} + N_2 C_{23} &= N_3(C_{31} + C_{32} + A_{32} + A_{31}) \\
N_1 C_{12} + N_3(C_{23} + A_{32}) &= N_2(C_{23} + C_{21} + A_{21}) \\
N_1 + N_2 + N_3 &= 1
\end{aligned}
\tag{3.15}
$$

The last line normalises of the populations in all of the levels to sum to
unity, so the populations of individual levels are given as a fraction per ion.
Clearly, these are three linear equations, and can be readily solved. However,
to illustrate the way in which such an atom can be used to determine the
physical parameters of temperature and density, it is interesting to consider
two special cases.

3.2.1 Low Density Limit; $E_{12} \sim E_{23}$

In this case, because of the low density, collisional de–excitations of the ex-
cited levels can be safely ignored $(C_{ij} \sim 0$ for $i > j)$. Also, because of the
increasing threshold energies to excite each level, $N_3 \ll N_2 \ll N_1$ so that
equation (3.15) can be reduced to:

$$
\begin{aligned}
N_1 C_{13} &= N_3(A_{32} + A_{31}) \\
N_1 C_{12} + N_3 A_{32} &= N_2 A_{21} \\
N_1 + N_2 + N_3 &= 1
\end{aligned}
\tag{3.16}
$$

hence, $N_3 = N_1 C_{13}/(A_{32} + A_{31})$ and $N_2 = N_1 C_{12}/A_{21}$. If we now form the
line intensity ratio for the $3 \rightarrow 2$ and $2 \rightarrow 1$ transitions, and substitute
equations (3.3) and (3.4) in the resulting expression we have:

$$
\begin{aligned}
\frac{F_{32}}{F_{12}} &= \frac{E_{32} A_{32} C_{13}}{E_{21}(A_{32} + A_{31}) C_{12}} \\
&= \frac{E_{32} A_{32} \Omega_{13}}{E_{21}(A_{32} + A_{31}) \Omega_{12}} \exp\left[\frac{-E_{23}}{kT}\right]
\end{aligned}
\tag{3.17}
$$
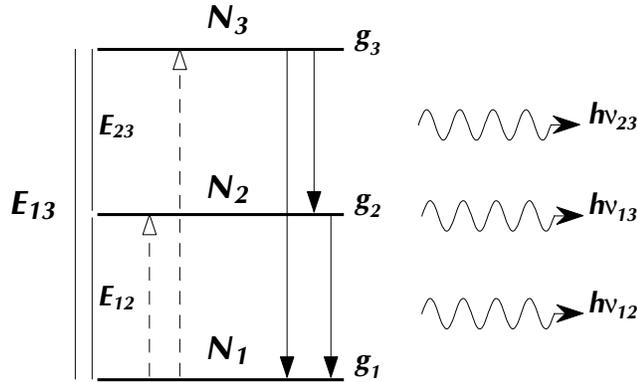
Fig. 3.3. A three–level atom with nearly equi–spaced excited states. Such a level configuration allows the ion to be used for temperature diagnostic purposes.

Because this line ratio varies as $\exp\left[-E_{23}/kT\right]$, it can be used to measure the electron temperature in the plasma.

Astrophysical examples of such temperature–sensitive emission line ratios are to be found in the forbidden lines of the p^2 and p^4 ions. These are typically excited by thermal electrons at nebular temperatures $\sim$10000K. An example of the temperature diagnostics is shown for the case of the [O III] ion in Figure (**??**). The reason we can use the three level approximation for ions such as these is that the characteristic resonance lines have much higher threshold energies, and are unimportant as cooling transitions at these temperatures and densities.

For these atoms, the ground term is a triplet, as a result of the spin–orbit interactions. However, the ground term splitting is small, so the atom can be approximated by a three–level system. In addition, the relative intensities of the individual transitions to the ground term can easily be worked out. Quantum mechanics shows that, apart from relativistic corrections;

$$A(^1\mathrm{D}_2 \to^3 \mathrm{P}_2) : A(^1\mathrm{D}_2 \to^3 \mathrm{P}_1) \approx 3 : 1 \tag{3.18}$$

and

$$A(^1\mathrm{S}_0 \to^3 \mathrm{P}_2) : A(^1\mathrm{S}_0 \to^3 \mathrm{P}_1) \approx 3 : 1. \tag{3.19}$$

In Table 3.1 we list the wavelengths of some important nebular lines in p^2 and p^2 ions. Note that in the p^4 ions the order of the ground term energy levels is reversed.

[OIII]λ5007,4363
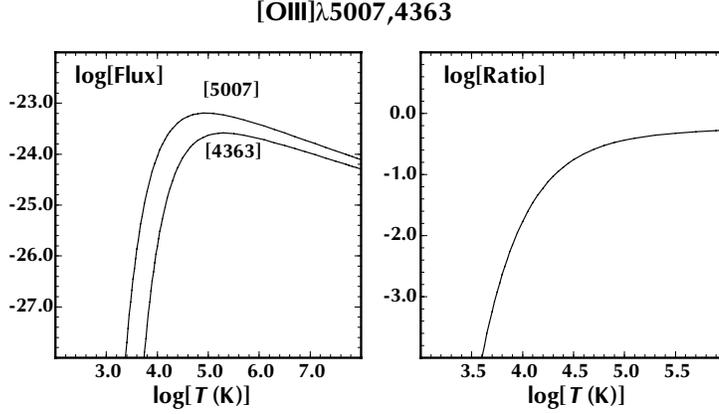
Fig. 3.4. The temperature dependence of the emissivity (left) and the ratio (right) of the two [O III] lines, λ 5007Å and λ 4363Å .

Table 3.1. Some important nebular lines (Å)

p^2 Ions	[N II]	[O III]	[Ne V]	[S III]
$^1S_0 \rightarrow {}^1D_2$	5755	4363	2974	6312
$^1D_2 \rightarrow {}^3P_2$	6583	5007	3426	9532
$^1D_2 \rightarrow {}^3P_1$	6548	4959	3346	9069
p^4 Ions	[O I]	[Ne III]	[Ar III]	
$^1S_0 \rightarrow {}^1D_2$	5577	3343	5192	
$^1D_2 \rightarrow {}^3D_2$	6300	3869	7136	
$^1D_2 \rightarrow {}^3P_1$	6363	3968	7751	

3.2.2 Ions in which $E_{23} \ll E_{12}$

For three–level ions configuration like that shown in figure (3.5), and in the low density limit, we can neglect collisionally induced transitions between the higher levels. Hence, $N_1C_{12} = N_2A_{21}$ and $N_1C_{13} = N_3A_{31}$. Therefore, in the low density case the line flux ratio is given by :

$$\frac{F_{32}}{F_{12}} = \frac{E_{32}A_{32}N_3}{E_{21}A_{21}N_2} = \frac{E_{31}C_{13}}{E_{21}C_{12}} = \frac{\Omega_{31}}{\Omega_{21}} \exp\left[\frac{-E_{23}}{kT}\right] \sim \frac{\Omega_{31}}{\Omega_{21}}. \qquad (3.20)$$

using the quantum mechanical sum rule for collision strengths equation (3.6) yields the result:
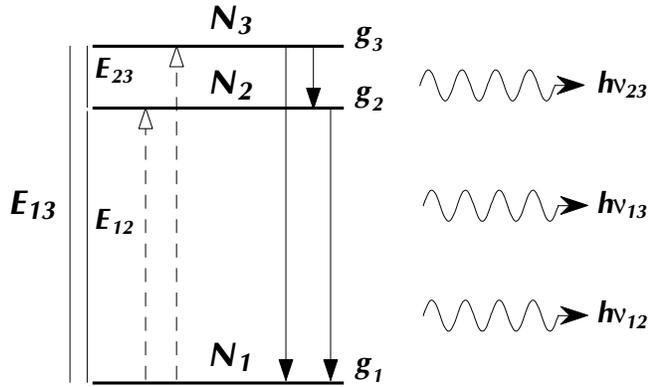
Fig. 3.5. An idealised three–level ion with a small energy separation between the excited states. Ions having configurations like this can be used for density diagnostic purposes,

$$\frac{F_{32}}{F_{12}} \sim \frac{\Omega_{31}}{\Omega_{21}} = \frac{g_3}{g_2}. \tag{3.21}$$

On the other hand, in the high density limit, the upper levels are populated according to their Boltzmann ratios, $N_3/N_2 = g_3/g_2$. Therefore, the line flux ratio:

$$\frac{F_{32}}{F_{12}} = \frac{E_{32}A_{32}N_3}{E_{32}A_{32}N_3} = \frac{A_{31}g_3}{A_{21}g_2}. \tag{3.22}$$

Note that in both cases, the small energy gap between the second and third level has the effect of supressing the temperature dependence, *i.e.* as $\Delta E/kT \ll 1$, $\exp[\Delta E/kT] \to 1$. These equations imply that the line ratios at low density and at high density both tend to a limit, and that these limits are different, provided that the transition probabilities are not the same. Thus lines of ions with configurations like this can be used as density diagnostics in the regime between the critical densities for de–excitation of each of the transitions. Equation (3.14) shows that the critical densities scale as $T^{1/2}$, so that the line intensity ratio in this intermediate range of density, where the line ratio is sensitive to the electron density, is only weakly dependent upon temperature.

The physical reason the line ratio is dependent on density can be understood if we think of the atom as behaving as two independent two–level atoms involving the second level and the ground state and the third level

[SII] λλ **6731,17**
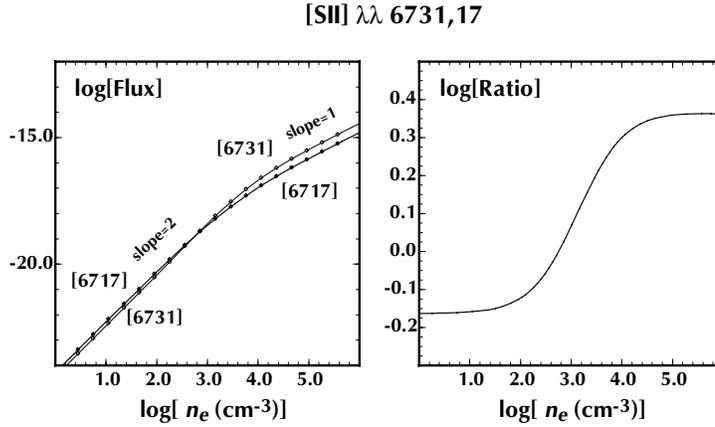
Fig. 3.6. The behaviour of the emissivity (left) and the line ratio (right) as a function of density in the [S II] doublet which provides useful density diagnostics.

Table 3.2. Some density sensitive nebula lines(Å)

p^2 Ions	[O II]	[S II]	[Ne IV]	[Ar IV]
$^2D_{3/2} \rightarrow {}^4S_{3/2}$	3726	6731	2423	4740
$^2D_{3/2} \rightarrow {}^4S_{3/2}$	3729	6716	2426	4711

and the ground state. For each line, the line intensity increases as the square of the density, according to equation (3.11), up to the critical density given by equation (3.14), after which it continues to increase proportional to the density according to equation (3.12). Since each level has a different critical density, the line ratio changes smoothly in the range between the two critical densities. This is illustrated in Fig. (3.6), which illustrates the specific case of the [S II] ion at an electron temperature of 10^4K.

Such transitions are found in the forbidden lines of the ground terms of the p^3 ions. For example:

These ions also have a higher ^{2}P term, which is well separated from the ^{2}D term in energy. Thus transitions between these terms form temperature sensitive line ratios, while transitions between the two lower terms are density sensitive. Thus, observations of both line ratios allows for a simultaneous solution of both temperature and density in the nebular zone containing these ions. An example of such a pair of line ratios is, in the [O II] ion the $\lambda\lambda(7318+7329)/(3726+3729)$Å and the $\lambda\lambda(3726)/(3729)$Å line ratio. An-
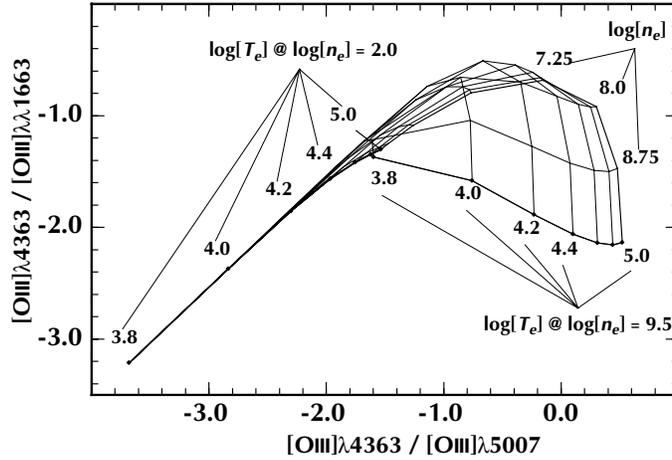
Fig. 3.7. The temperature and density dependence of the O III forbidden line ratio $\lambda\lambda 4363/5007$Å and the forbidden to intercombination line ratio $\lambda\lambda 4363/1663$Å. At low densities both are temperature diagnostics, but at high enough densities, both density and temperature can be determined. This pair of ratios would be a useful diagnostic in objects such as young nova shells, and in active galactic nuclear emission line regions.

other example is provided by the [S II] $\lambda\lambda(4069+4076))/(6716+6731)$Å and the $\lambda\lambda(6731)/(6716)$Å line ratio.

In practice, almost any two line ratios for a given ion can be used for such diagnostics. At successively higher densities, the ratio of a intercombination to a forbidden line, the ratio of a resonance to an intercombination or the even ratio of two resonance lines may be used. An example of the use of this is shown for the case of the O III ion in Figure (3.7). Here we plot the ratio which is usually used as a temperature sensitive line ratio at low density, the O III(λ 4363/λ 5007Å) line ratio against the ratio of the intercombination and forbidden line, the O III($\lambda 4363/\lambda\lambda(1660+1666)$Å)line ratio (which is also sensitive mainly to temperature at low densities). See Figure (2.4) for the terms involved.

Note the way in which each of these ratios becomes sensitive to density in turn, once the critical density for one of the lines involved in the line ratio is exceeded. For forbidden lines, the range of density sensitivity is typically 100 $– 10^7$ cm^{-3}, while intercombination lines extend the range up to $\sim 10^{10}$ cm^{-3}, and resonance lines could, in principle, be used up to the point of transition to full LTE conditions. However, in practice, this is not usually possible, since

the radiative line transfer problems and radiative pumping of levels will tend to dominate.

3.2.3 Infrared Line Diagnostics

In dense regions of rapid star formation, or in regions close to the centres of galaxies, the visible light is often blocked from view by the surrounding dust. In order to probe the conditions of such regions, we need to observe in the far–infrared, where the dust obscuration is low, allowing us to probe these dense regions. In addition, the dust itself displays a rich emission spectrum at these wavelengths, allowing us to identify the nature, composition, and size distribution of the dust component itself. The advent of the *Infrared Astronomy Satellite* (IRAS) and the *Infrared Space Observatory* (ISO) opened up this far–infrared window of the electromagnetic spectrum to detailed study. It is necessary to observe in space because at most ground–based sites, the region of the spectrum covering roughly 15-250μm is blocked from observation by atmospheric absorption, mainly caused by water vapour.

The far–IR region of the spectrum contains most of the emission lines which are responsible for cooling plasmas with temperatures of between 100 – 3000 K, but these lines are still important in even hotter plasmas. The principal ions which give rise to lines in this region are the p^2 and p^4 ions, since these have the multiplet ground terms in which hyperfine transitions may occur. Because the transition probabilities are low in these transitions, the critical densities are also quite low (unless the species is highly ionised), so that ratios of lines from the same ion form useful density diagnostics. However, the mean thermal energy of the electrons or ions is usually appreciably higher than the excitation energies of the upper states, so nothing can usually be gleaned about the temperature. More useful is the fact that the emissivity in any line which is not affected by collisional de–excitation is simply proportional to the ionic abundance, thus these lines be used as abundance diagnostics, and ratios of lines of different ionisation may also be used to measure the excitation of the plasma. In Table (3.2), we list some of the most important lines detected with ISO. Here the critical density is given in terms of the collisions with atomic hydrogen. Collisions with molecular hydrogen may also be important for IR transitions involving un–ionised atomic species.

3.3 The General Multi–Level Atom

It is easy to generalise the equations of statistical equilibrium given in equation (3.15) up to an arbitrary number of levels. In equilibrium, the rate of collisional and radiative population of any level is matched by the collisional and radiative depopulation rates of that same level. When combined with the

Table 3.3. Important lines detected with ISO

Species	Transition	$\lambda(\mu m)$	$n_{\text{crit.}}$ (cm^{-3})	IP(eV)
[C II]	$^2P^0_{3/2} - {}^2P^0_{1/2}$	157.74	3×10^3	11.26
[O I]	$^3P_0 - {}^3P_1$	145.50	9×10^4	...
[O I]	$^3P_1 - {}^3P_2$	63.18	5×10^5	...
[O III]	$^3P_1 - {}^3P_0$	88.36	4×10^3	35.12
[O III]	$^3P_2 - {}^3P_1$	51.82	5×10^2	35.12
[O IV]	$^2P_{3/2} - {}^2P_{1/2}$	25.87	1×10^4	54.93
[N II]	$^3P_1 - {}^3P_0$	203.5	5×10^1	14.53
[N II]	$^3P_2 - {}^3P_1$	121.9	3×10^2	14.53
[N III]	$^2P_{3/2} - {}^2P_{1/2}$	57.32	3×10^3	29.60
[N IV]	$^3P_1 - {}^3P_2$	...	1×10^6	47.45
[Ne II]	$^3P_{1/2} - {}^3P_{3/2}$	12.81	5×10^5	21.56
[Ne III]	$^3P_0 - {}^3P_1$	36.02	4×10^4	40.96
[Ne III]	$^3P_1 - {}^3P_2$	15.55	3×10^5	40.96
[Ne v]	$^3P_1 - {}^3P_0$	24.28	5×10^4	126.2
[Ne v]	$^3P_2 - {}^3P_1$	14.33	4×10^5	126.2
[S III]	$^3P_1 - {}^3P_0$	33.48	2×10^3	23.33
[S III]	$^3P_2 - {}^3P_1$	18.71	2×10^4	23.33
[S IV]	$^2P_{3/2} - {}^2P_{1/2}$	10.51	6×10^4	34.83
[Si II]	$^2P^0_{3/2} - {}^2P^0_{1/2}$	34.81	3×10^5	8.15
[Ar II]	$^2P_{1/2} - {}^2P_{3/2}$	6.99	2×10^5	6.99
[Ar III]	$^3P_1 - {}^3P_2$	8.99	3×10^5	27.63
[Ar III]	$^3P_0 - {}^3P_1$	21.8	3×10^4	27.63
[Fe II]	$^6D_{7/2} - {}^6D_{9/2}$	25.99	2×10^6	16.18
[Fe II]	$^6D_{5/2} - {}^6D_{7/2}$	35.35	3×10^6	16.18

population normalisation equation (the sum of the populations of all levels must add up to the total number of ions, we have a linear set of simultaneous equations which may be solved in the standard way. Formally the equations of statistical equilibrium for all levels j and the population normalisation equation can be written:

$$\sum_{J \neq j}^{\infty} N_J C_{Jj} + \sum_{J>j}^{\infty} N_J A_{Jj} - N_j \left(\sum_{J \neq j}^{\infty} C_{jJ} + \sum_{J<j}^{\infty} A_{jJ} \right) = 0$$

$$\sum_{J=1}^{\infty} N_J = 1. \qquad (3.23)$$

This can be expressed in a more convenient form by splitting the collisional excitation (E) rates from the collisional de–excitation (D) rates as:

$$\sum_{J}^{\infty} N_J (C^D_{Jj} + C^E_{Jj} + A_{Jj}) - N_j \sum_{J}^{\infty} \left(C^D_{jJ} + C^E_{jJ} + A_{jJ} \right) = 0$$

$$\sum_{J=1}^{\infty} N_J = 1, \quad (3.24)$$

with:

$$A_{Jj} = 0; \qquad C_{jJ}^D = 0; \qquad J \geq j$$
$$C_{jJ}^E = 0; \qquad 1 \leq J \leq j.$$

Since the population of the higher levels is usually much lower than that of the lower levels, a rather stable way of solving these equations is by diagonalising the matrix, followed by back substitution to solve for individual level populations, starting with the most excited. After equations (3.24) have been solved to give the population of each level, the rate of cooling in each transition, F_{Jj}, can be then calculated from the level populations and the ionic abundance (by number, relative to hydrogen N_H), η_i,

$$F_{Jj} = \eta_i N_H N_J E_{jJ} A_{Jj}. \qquad (3.25)$$

At low density, the populations N_J are proportional to the square of the density, or more precisely, the product of the electron density and the number density of the ion. So we can introduce the idea of a total emission line *cooling function* by summing the fluxes in all the radiative transitions for all the atoms and ions in the plasma;

$$\Lambda_{line} = \left(\frac{1}{n_H n_e}\right) \sum_{A,I} Z_{A,I} \sum_{J=1}^{\infty} F_{Jj}. \qquad (3.26)$$

In the low density limit, the density dependence of the emission is removed in this expression, and Λ_{line} is effectively a function of temperature only. The implicit assumption here is that the emission flux produced in each line is free to escape from the region of diffuse plasma where it is generated.

Exercise 3.3.1. An atom has three fine structure states, ^{3}P, ^{1}S$_0$ (Excitation Energy 0.5eV) and ^{1}D$_2$ (Excitation Energy 1.2 eV). From the ground term, the total collision strength to both excited levels is 1.0. Three transition probabilities are $A(^1$S$_0$ - ^{3}P$_1) = 5$ s^{-1} , $A(^1$D$_2$ - ^{3}P$_1) = 20$ s^{-1} and $A(^1$D$_2$ - ^{1}S$_1) = 10$ s^{-1}.

1. What are the transition probabilities $A(^1$S$_0$ - ^{3}P$_2)$ and $A(^1$D$_2$ - ^{3}P$_2)$?
2. Assuming that the ground term splitting is negligible, what are the wavelengths of the forbidden lines produced by the atom?
3. In this case, what is the critical density for each forbidden line?
4. Plot the flux ratio F(^{1}D$_2$ - ^{3}P$_{1,2}$) / F(^{1}S$_0$ - ^{3}P$_{1,2}$) as a function of temperature for low densities, and at the limit of high densities.

Exercise 3.3.2. Here is a set of atomic data for a five–level atom: J-Quantum numbers of Levels 1 through 5: 3/2, 5/2, 3/2, 3/2, 1/2

1. What are the statistical weights of each level?
2. What are the wavelengths (nm) of the transitions produced by this ion?
   (1%)
3. Using any technique (Maple, C–program, Mathematica etc...), plot the
   intensity ratio of the (2 - 1) and the (3 - 1) transitions as a function of
   electron density, log (n_e), in the range $1 \leq \log (n_e) \leq 5$. Please show
   your working.

Energy Level Matrix (ergs):

Level	1	2	3	4	5
1	0	5.32589E-12	5.32976E-12	8.03878E-12	8.03900E-12
2		0	3.87000E-15	2.71289E-12	2.71311E-12
3			0	2.70902E-12	2.70924E-12
4				0	2.2E-16
5					0

Transition Probability Matrix (s^{-1}):

Level	1	2	3	4	5
1	0	3.82E-05	1.65E-04	5.64E-02	2.32E-02
2		0	1.20E-07	1.17E-01	6.15E-02
3			0	6.14E-02	1.02E-01
4				0	0
5					0

Collision Strength Matrix :

Level	1	2	3	4	5
1	0	8.01E-01	5.34E-01	2.70E-01	1.35E-01
2	8.01E-01	0	1.17E+00	7.30E-01	2.95E-01
3	5.34E-01	1.17E+00	0	4.08E-01	2.75E-01
4	2.70E-01	4.08E-01	7.30E-01	0	2.87E-01
5	1.35E-01	2.95E-01	2.75E-01	2.87E-01	0

4. Line Transfer Effects

"There are no such things as applied sciences,
only applications of science"
— Louis Pasteur

4.1 Resonance Line Transfer

The discussion of the previous chapter has established the way in which emission lines may be produced by collisional excitation locally within a nebula. However, whether the observer can see the emission depends upon whether the nebula is optically thick or optically thin to the escape of this radiation. This in turn depends upon the *optical depth* of the nebula at the frequency, ν, considered; τ_ν. For transmission of light through an absorbing screen of material, the optical depth is defined by the factor by which the intensity of the radiation has been reduced, $I(\nu)/I_0(\nu) = \exp[-\tau_\nu]$. The optical depth is the integral along the line of sight of the *linear absorption coefficient* (cm^{-1}) $\kappa(s,\nu)$; $\tau_\nu = \int \kappa(s,\nu)ds$. However a more useful measure is the mass absorption coefficient (cm^{2}g^{-1}), κ_ν, which measures the effective cross section contributed per unit mass of matter at this frequency, $\kappa_\nu = \kappa(s,\nu)/\rho(s)$. This is also called the *opacity* of matter. Since matter can either scatter light (re–radiation at the same frequency), or absorb it one frequency to re–radiate at another, we have to recognise that the total opacity is the sum of these two contributions $\kappa_\nu = \kappa_\nu(scat) + \kappa_\nu(abs)$. Absorption is produced by either dust, nebular continuum processes, or by fluorescence (degradation of the original photon into two or more photons, which occurs when the excited state of the atom has more than one permitted decay route). Scattering can be produced by free electrons, or by resonant scattering in lines.

In the case of a scattering screen, a photon in the beam of light from a distant object is first removed from the beam, and then re–radiated in all directions, so is effectively lost from the beam. Thus, in resonance line absorption by the ISM, we see a series of dark lines superimposed on the intrinsic continuum spectrum of the distant object (star, galaxy or whatever). Let us first consider this process in a little more detail.

4.1.1 Resonance Line Absorption by Heavy Elements

Consider an excited state, $j = 2$, and the corresponding resonance transition to the ground state, $j = 1$. The width in energy (or frequency) of an excited state is not a perfect delta function. This is because the lifetime Δt, of the excited state is finite, determined by the transition probablility back down to the ground state, A_{21}; given by equation (??) or (2.5): $\Delta t = 1/A_{21}$. This finite lifetime implies finite energy width by the Heisenberg uncertainty principle of quantum mechanics; $\Delta E.\Delta t \geq \hbar$. Thus, this energy uncertainty translates to a natural line width in terms of frequency (full width half maximum; FWHM), $\Delta \nu_N = A_{21}/2\pi$. Since the frequency dependence of the absorption profile can be represented as the Fourier transform of an exponentially decaying harmonic oscillator, the effective absorption coefficient at frequency ν has a natural line shape which is Lorentzian, centered at the frequency of the transition ν_0 and having a FWHM $\Delta \nu$. In terms of the population of the ground level, N_1, and the absorption oscillator strength f_{12}, the cross section is given by:

$$\sigma(\nu) = \frac{\pi \mathbf{e}^2}{m_e c} f_{12} N_1 \left[1 - \frac{g_1 N_2}{g_2 N_1} \right] \frac{\Delta \nu / 2\pi}{(\nu - \nu_0)^2 + (\Delta \nu / 2)^2} \qquad (4.1)$$

Here, the term in square brackets contains the factor which corrects for the stimulated emission, and the frequency function has been normalised so that its integral over the full line profile is unity. This function declines as a power law with frequency for frequencies which lie far away from the line core. These are often referred to as the *damping wings* of the profile. If we neglect the correction for stimulated emission, and assume that at the low density limit all of the atoms are in their ground state, then at the line centre the cross section per atom is $\sigma(\nu_0) = 2\mathbf{e}^2 (m_e c)^{-1} f_{1j} \Delta \nu_N^{-1} \sim 0.0169 f_{1j} \Delta \nu^{-1} \text{ cm}^2$.

In most cases, near the line core, the line broadening is not dominated by the natural width, but by the Doppler broadening caused by the thermal motions of the atoms along the line of sight, v_x. The Doppler shift in frequency produced by this velocity relative to the line center is given by $(\nu - \nu_0) = \nu_0 v_x / c$. At a given ionic temperature T, the fraction of ions, with mass M, in the velocity range v_x to $v_x + dv_x$ is given by the one dimensional Maxwell distribution:

$$dN(v_x) = \left(\frac{M}{2\pi KT} \right)^{1/2} \exp \left[\frac{-M v_x^2}{2kT} \right] dv_x. \qquad (4.2)$$

This will produce an absorption cross section profile which is Gaussian:

$$\sigma(\nu) = \frac{\pi \mathbf{e}^2}{m_e c} f_{12} N_1 \left[1 - \frac{g_1 N_2}{g_2 N_1} \right] \left(\frac{M}{2\pi KT} \right)^{1/2} \exp \left[\frac{-M c^2 (\nu - \nu_0)^2}{2kT \nu_0^2} \right]. \qquad (4.3)$$

In this case the line width of the Doppler profile, $\Delta \nu_D$, (FWHM) is given by $\Delta \nu_D = 2(\ln 2)^{1/2} (2kT/M)^{1/2} (\nu_0/c)$. Again, neglect the correction for stimulated emission, and assume that at the low density limit all of the atoms are

in their ground state, then at the line centre, the cross section per atom is $\sigma(\nu_0) = 2(\pi \ln 2)^{1/2} \mathbf{e}^2 (m_e c)^{-1} f_{1j} \Delta \nu_D^{-1} \sim 0.0249 f_{1j} \Delta \nu_D^{-1}$ cm^2. In addition to these natural and thermally–broadened line profiles, the absorption line will also be broadened by the internal turbulence of the gas cloud in which it is formed, and shifted to higher or lower frequencies by the Doppler shift due to bulk motion of the cloud with respect to the observer.

What is the effect of these absorbers on the light of a continuum source of intensity I_0 located behind the cloud? At any frequency, the fraction of light which gets through is determined by the optical depth at that frequency, $\tau(\nu) = N\sigma(\nu)$, where N is the total column density of absorbing atoms through the cloud. Thus:

$$I(\nu) = I_0 \exp\left[-N\sigma(\nu)\right] \tag{4.4}$$

It is convenient to define an *equivalent width*, W, for the line, which measures the net flux removed from the incident beam by the absorption line. This quantity is useful because it can be measured even when the spectrograph used to observe it does not have the resolution needed to resolve the details of the actual line profile. By the definition of equivalent width, the total flux removed from the beam by the line is $I_0 W$. We can also write this product as:

$$I_0 W = I_0 \int d\nu - \int I(\nu) d\nu \tag{4.5}$$

that is, the flux that was in the beam before absorption less the flux that remains in the beam after absorption. However, from this equation, and equation (4.4), it is clear that we can write W as:

$$W = \int \left(1 - \exp\left[-N\sigma(\nu)\right]\right) d\nu \tag{4.6}$$

When the optical depth of the cloud is small at all frequencies, $\tau(\nu) = N\sigma(\nu) \ll 1$, it follows that $W = N \int \sigma(\nu) d\nu$ so that the equivalent width is directly proportional to the number of atoms on the line of sight. If this condition is satisfied, then we refer to the line as lying on the *"linear"* portion of its *curve of growth*. From observational usage, the equivalent width is often presented in terms of wavelength rather than frequency. In this convention, a very convenient form for the column density in the linear portion of the curve of growth is:

$$\left[N/\text{cm}^{-2}\right] = 1.13 \times 10^{17} f_{12}^{-1} \left[\lambda/\text{Å}\right]^{-2} \left[W_\lambda/\text{mÅ}\right] \tag{4.7}$$

This equation is extensively used in observational studies.

When the line becomes optically thick in its core, then the residual flux in these regions is very small, and so the equivalent width can only increase as a result of absorption in the Doppler wings of the profile in equation (4.3). In this case, the effective width of the line core is determined by the condition

that $\tau(\nu) = N\sigma(\nu) \sim 1$. Because this effective width grows only rather slowly, there is only a weak dependence of equivalent width on column density, $W \propto (c + \ln N)^{1/2}$. This is referred to as the *"flat"* portion of the curve of growth. In this region very little can be inferred about the column density at low resolution, and any meaningful analysis requires observations made at very high resolution and excellent signal to noise to study the detailed line profile in the wings. This gives the apparent column density profiles, from which true column densities may be derived provided that the smearing produced by the instrumental response can be adequately modelled and corrected for.

Finally, when the column density is very high indeed, the relationship between the equivalent width and the column density is dominated by the damping wings of the Lorentzian line profile of equation (4.1). In this *"damping"* portion of the curve of growth, the equivalent width grows rather more quickly once more, with a square root dependence, $W \propto N^{1/2}$. This is conveniently rendered for the observationalist as:

$$\left[N/\mathrm{cm}^{-2} \right] = 1.7 \times 10^{32} f_{12}^{-1} \Delta\nu^{-1} \left[\lambda/\text{Å} \right]^{-4} \left[W_\lambda/\mathrm{mÅ} \right]^2. \qquad (4.8)$$

In the interstellar medium, column densities this large are usually only seen in the case of the Lyα line or in the Mg II doublet $\lambda\lambda$ 2796,2803Å.

If we want to accurately derive the column density of a particular ion, it is best if a fairly large number of transitions with widely different gf values are observed, so that absorption lines in the linear, flat and damping portions of the curve of growth are all observed. This is possible for lines formed in the atmospheres of stars, and here the damping portion of the curve of growth can also be used to measure the pressure broadening of the natural line width caused by collisions between atoms effectively shortening the lifetime of the excited state. However, in the interstellar and intergalactic medium, we can rarely observe transitions of a particular ion covering the full curve growth.

4.1.2 Absorption Line Studies of the ISM

The techniques described in the previous section have been widely and extensively applied to studies of the ISM, and indeed, to studies of the IGM as well. The virtue of the technique is the sensitivity to very low column densities of gas. Consider, for example the case of the Zn II line at $\lambda 2026$Å, which has been used to study the chemical evolution of the IGM in lines of sight to high redshift QSOs. This line has $f_{12} = 0.489$, so from equation (4.7), a line with an equivalent width of 100mÅ (easy to measure with modern instrumentation on a large telescope) is produced by a column density of only $\sim 6 \times 10^{12}\,\mathrm{cm}^{-2}$. If Zn II is the main ionisation stage present, then this corresponds to a hydrogen column density of $\sim 3 \times 10^{17}\,\mathrm{cm}^{-2}$. Such a column density would be very hard to detect if we were relying upon the emission measure of the hydrogen gas, and any emission in this zinc line would lie many orders of magnitude below the sensitivity limits of our instruments.

Relatively few lines of atomic or molecular species are accessible from ground–based observation. However, the number of species which are observable increases strongly as we penetrate further into the ultraviolet. The wavelength range between 1200 and 3200Å was opened up with the advent of the IUE satellite, and has been exploited by the much greater capabilities of the GHRS instrument on the Hubble Space Telescope (HST). However, the region shortward of Lyα has been little studied between the 1970s, when the Copernicus satellite first opened up this region to study, and the late 1990s with the launching of the FUSE (Far Ultraviolet Spectroscopic Explorer) satellite. These lines can be resolved into many cloud components along the line of sight, and are observed both on their linear portions of their curves of growth, and into saturation. Figure (4.1) shows some of the remarkable data that has been obtained with the Hubble Space Telescope in a line of sight to a reddened halo star in the Galaxy. Here we can see lines in the linear portion of their curve of growth, and others that have gone well into saturation on the flat part of the curve of growth. Table (4.1.2) lists the species which have so far been detected in the ISM.

Table 4.1. Interstellar absorption lines detected in the ISM.

Optical	$\lambda > 3200$Å
Atoms:	Li I, Ca I, K I, Fe I, Ca II, Ti II
Molecules:	C_2, CO, CO^+, N_2, NO, NO^+, OH, CN^+, CS, SiO, MgH^+, HCl, CH_2, H_2O
UV	1200Å$< \lambda < 3200$Å
Atoms:	H I, D I, C I, C II, C IV, N I, N V, O I, Mg I, Mg II
	Al II, Al III, Si I, Si II, Si III, Si IV, P I, P II, P III, S I,
	S II, S III, Cl I, Cr II, Mn II, Fe II, Co II, Ni II, Cu II,
	Zn II, Ga II, Ge II, As II, Se II, Kr II, Sn II, Tl II, Pb II
Molecules:	H_2, ^{12}CO, ^{13}CO, $C^{17}O$, $C^{18}O$, C_2, HCl

One of the most startling successes of these observations has been the very direct evidence that has been gleaned about the nature and chemical composition of interstellar dust. When the gas phase abundances derived for a gas cloud lying along the line of sight to a fairly heavily reddened nearby star, such as ζOph, are compared with the abundances derived for, say, the sun, it is found that the heavy non–volatile elements are very strongly *depleted* in their relative abundances. The inescapable conclusion is that the elements missing from the gas phase exist in the solid form, as interstellar dust grains. A very striking correlation between the condensation temperature and the degree of depletion is found, *see* Figure (4.2).
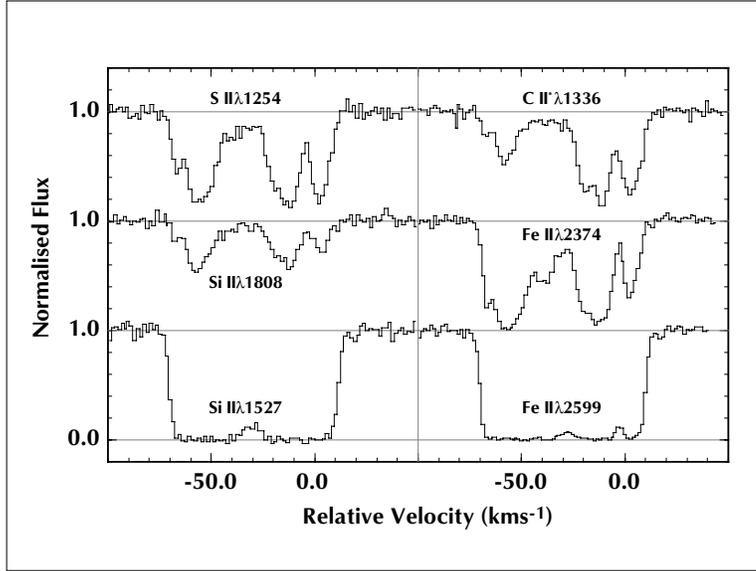
Fig. 4.1. Interstellar line profiles observed towards the reddened halo star HD 93521. Many velocity components are evident in this montage which covers both weak and strong lines. Analyses of such data give the physical conditions and chemical abundances in the different clouds (after Spitzer & Fitzpatrick, 1993).

Along other lines of sight, or even in other cloud components along the same line of sight, we see different patterns of depletion. These are connected with the temperature, density and thermal history of the ISM within them, and collectively they give insight into the way in which energetic processes in the ISM, such as shocks, or heating by radiation fields, can destroy the interstellar dust.

The ratio of deuterium to hydrogen in the cosmos is one of the key observational parameters which probe the nature of the Big Bang. This is because deuterium is made in the Big Bang nucleosynthesis, and the amount that is made is very sensitive to the cosmological parameters. The deuterium $Ly\alpha$ line lies 0.331Å to the shortward of the hydrogen $Ly\alpha$ line, thanks to the isotope shift which is due the difference in the reduced mass of the electron in deuterium as compared with hydrogen. Provided that the absorption line due to deuterium is strong enough to be detected in the wings of the hydrogen $Ly\alpha$ line, then the deuterium abundance can be accurately determined because the deuterium line lies on the linear section of the curve of growth, when equation (4.7) applies; whereas the hydrogen line is firmly in the damping portion of its curve of growth, for which equation (4.6) applies. This is evident from Figure (4.3). The GHRS observations for Capella yield (D/H) $= 1.60 \times 10^{-5}$ $(\pm 0.09 \times 10^{-5})$; $(+0.05; -0.10 \times 10^{-5})$ (Linsky *et al.* 1995),
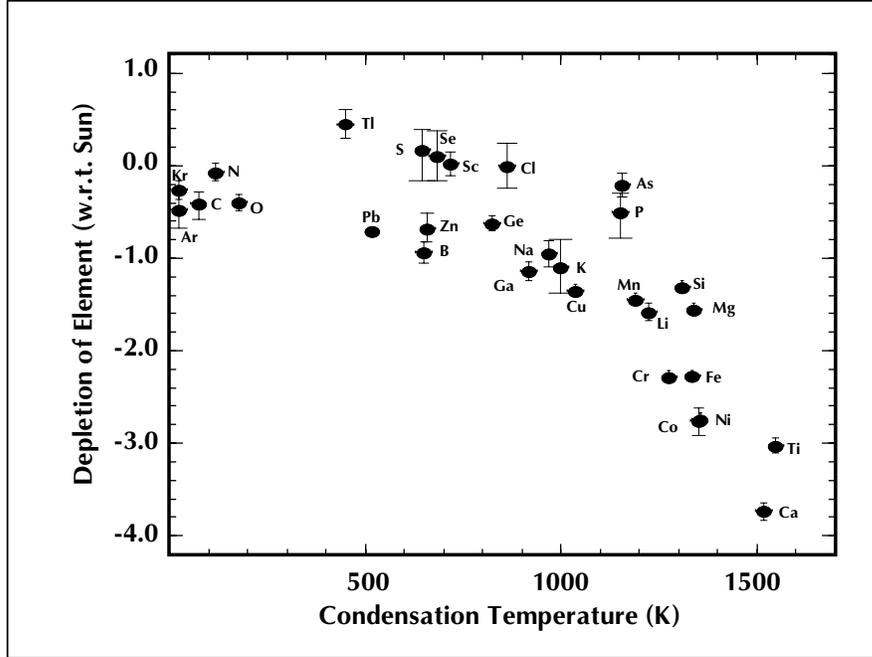
Fig. 4.2. The depletion of the gas–phase abundances of various elements, plotted as a function of their condensation temperatures for dust formation. This is derived from resonance line absorption measurements made along the line of sight to ζ Oph, which passes through a dense cool interstellar cloud (after Savage & Sembach, 1996).

which is currently one of the most accurate value yet determined for this ratio. During the process of nucleosynthesis, deuterium is easily destroyed in stars, measured interstellar abundances only provide an upper limit to the primordial deuterium abundance. Nonetheless, this is still sufficient to place a strong upper limit on the local baryon density of $\Omega_B h_{75} \lesssim 0.055$; where Ω_B is the ratio of the local baryon density to the critical density needed to close the universe, and h_{75} is the Hubble Constant in units of $75 \, \mathrm{km \, s^{-1} Mpc^{-1}}$. Even tighter limits can in principle be obtained by using lines of sight to distant QSOs. These probe the intergalactic medium present in the early universe, in which very little of the deuterium should have been destroyed.

In these distant reaches of space, a much greater proportion of the gas is located in clouds between the galaxies. Deep observations with large telescopes have revealed these from the Lyα absorption line which they produce. These enable us to detect H column densities as small as $10^{14} \, \mathrm{cm^{-2}}$. In the denser clouds, absorption lines of heavier elements such as Fe or Zn are also detected, which allow us to investigate the chemical evolution of the intergalactic medium (IGM) of the early universe.
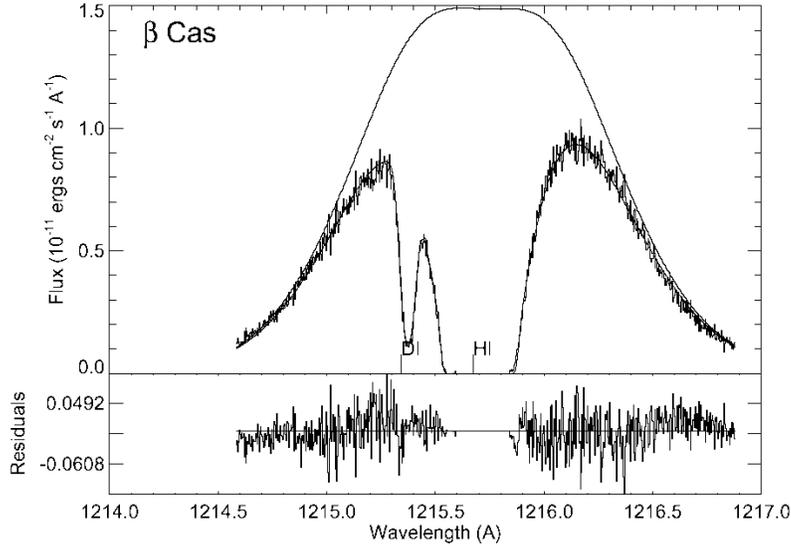
Fig. 4.3. The Lyα line profile in β Cas measured using the GHRS instrument on HST. The intrinsic line profile of the star is shown as a smooth curve, as is the model fit including absorption. The D/H ratio can be accurately measured because the deuterium line is detected in its linear portion of the curve of growth, while the H I line is in the damping portion (after Dring *et al.* 1997).

Finally, absorption line studies have been used to probe the hot phase of gas, and the nature of the diffuse plasma in the halo of our galaxy and in the Magellanic Clouds, using such ions as Si IV, C IV, and N V. In the local ISM this hot phase was also studied by the Copernicus satellite in the O VI ion. With the advent of the FUSE satellite, we will obtain a much deeper insight into this phase, and, linking these data with observations in the soft X–ray region of the spectrum, we should be able to finally understand the energy balance of hot galactic halos. Some of these applications will be discussed in more detail in section (7.6).

4.1.3 Line Transfer in Emission Resonance Lines

When the plasma is ionised, and hot enough that electron excitation becomes important, the equation of transfer in the line has to be modified to include the local emissivity, j_ν, of the plasma in the resonance line:

$$\frac{dI(\nu)}{ds} = -\kappa_\nu I(\nu) + j_\nu \tag{4.9}$$

Now, suppose that we have a slab of emitting material illuminated by some background source with intensity $I_0(\nu)$, then, substituting for the optical

depth $d\tau(\nu) = \kappa_\nu ds$, we can integrate equation (4.9) through the slab (of total optical depth $\tau_T(\nu)$) to give:

$$I(\nu) = I_0(\nu) \exp[-\tau_T(\nu)] + \int_0^{\tau_T(\nu)} \frac{j_\nu}{\kappa_\nu} \exp[\tau_T(\nu) - \tau(\nu)]d\tau(\nu) \qquad (4.10)$$

The first term represents the attenuated initial radiation field. Thus, when the line optical depth is much greater than unity, the emergent intensity is governed by the ratio j_ν/κ_ν in the cloud. From equation (4.9), in a nebula which is extremely optically thick in the line, so that $dI(\nu)/ds = 0$, the equilibrium intensity of the radiation field in the line is $I(\nu) = j_\nu/\kappa_\nu$. This radiation field represents a local thermal equilibrium, which, since it does not depend on any external factors, must be therefore be a function of temperature only. Common sense suggests that when this condition is satisfied, we have LTE, at least as far as the line is concerned. We can check this by writing down this condition of balance for the radiation energy density, that the emission rate in the line is equal to the absorption rate (corrected for the effect of stimulated emission):

$$\frac{4\pi I(\nu)}{c}(n_1 B_{12} - n_2 B_{21}) = n_2 A_{21} \qquad (4.11)$$

where B_{12} and B_{21} are the Einstein coefficients of absorption and stimulated emission, respectively. However, we have these from equations (**??**) and (**??**), *viz.* $g_1 B_{12} = g_2 B_{21}$ and $B_{21} = \left(c^3/8\pi h\nu^3\right) A_{21}$. Additionally, we can substitute the Boltzmann equation (3.2) for the populations in the ground and the excited state, since LTE applies between the levels. Substituting all of these in equation (4.11) we find the line intensity is in fact the Planck value appropriate to the plasma temperature:

$$I(\nu) = \frac{2h\nu^3}{c^2 \left(\exp\left[\frac{h\nu}{kT}\right] - 1\right)} = B(\nu, T) \qquad (4.12)$$

This means that, when the line is extremely optically thick, the radiation field is limited by its black–body value in the line, and that (from equation 4.10) Kirchhoff's law applies: $B(\nu, T) = j_\nu/\kappa_\nu$. At the low photon energies appropriate to radio wavelengths, the Rayleigh–Jeans limit of the black–body law applies $B(\nu, T) \sim 2KT\nu^2/c^2$, so we can re–write equation (4.10) in terms of the *brightness temperature* of the source. Brightness temperature is defined as the temperature that a black body would have to be at in order to give the same intensity $I(\nu)$ in the frequency range as is observed. This gives:

$$T_B = T_{B,0} \exp[-\tau_T(\nu)] + T\left(1 - \exp[-\tau_T(\nu)]\right) \qquad (4.13)$$

In the limit of high line optical depth, the fact that the photons observed arise in a very thin surface near the front edge of the cloud means that we can think of the cloud as possessing a *photosphere* in the line. As long as the line

remains optically thick, the emitted flux is limited at its black–body value, so that, provided the cloud is spatially resolved by the telescope which observes it, $T_B = T$ (minus 2.7K to correct for the cosmic microwave background radiation), which enables us to directly measure the cloud temperature at the photophere).

Over the fully optically–thick region of the observed line profile the emergent flux is constant at the black–body value. The existence of an effective photosphere in the line limits the total flux that can be emitted in the line; $F_{\text{line}} = A\Delta\nu B(\nu, T)$, where A is the emitting area, and $\Delta\nu$ is the line width. This also means that optically–thick lines cannot be used to probe the details of cloud structure inside the effective photosphere. Such a situation is encountered fairly frequently in the case of radio–frequency molecular transitions. The most famous (or infamous) example is the ^{12}CO(1-0) transition at $115.271\,$GHz – a key coolant in cool molecular clouds.

Because the line flux in an optically–thick line is independent of the abundance of the species which produce the line, we cannot use these lines to infer molecular ratios (such as the ^{12}CO/H$_2$ ratio, as has been frequently attempted in the literature, *c.f.* Section 13.1.4). Instead, we must find species which are optically thin, for which a cooling function such as equation (3.11) applies. In CO, such a line would be the ^{13}CO(1-0) $110.201\,$GHz transition, or perhaps a higher transition such as ^{13}CO(2-1) at $220.399\,$GHz. However, such lines have much lower surface brightnesses and are therefore much harder to observe.

Let us now look at the problem of resonant line scattering. If the photons are produced in an (optically–thick) region of radius R, the line will, in general, be scattered before it escapes. Generally, for the line transfer problem, we adopt the *Sobolev Approximation*. In this we assume that each line re–emission process is independent of the line absorption process which preceded it. Thus, regardless of where in the line profile the absorption occurred, the re–emitted photon will be redistributed in frequency according to the Doppler line profile, with a probability amplitude that is the same as the (normalised) Doppler profile. Thus, the Sobolev approximation is tantamount to assuming that the scattering redistributes the photon into the original Doppler profile. In this case, most of the photons escaping the nebula can only do so by being scattered into the wings of the profile where the optical depth is low.

The *escape probability* depends on the details of the geometry and dynamics of the region. A simple example is that of an (unexpanding) spherical nebula or radius R, and central optical depth $\tau_T(\nu)$.

If j_ν is the line emissivity, then the emergent intensity along the ray shown is:

$$
I(\theta, \nu) = \int_0^{2\tau_T(\nu)\cos\theta} \frac{j_\nu}{\kappa_\nu} \exp[-\tau(\nu)]d\tau(\nu)
$$

$$
= \frac{j_\nu}{\kappa_\nu}\left(1 - \exp[-2\tau_T(\nu)\cos\theta]\right) \tag{4.14}
$$
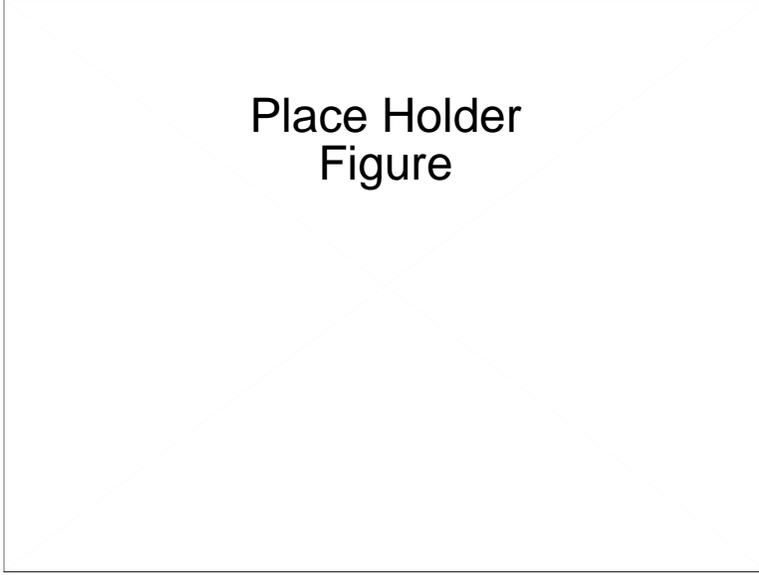
Fig. 4.4. The escape probability for a sphere, needs a better caption.

Hence, the total outward flux per unit area and per unit time is given by integrating over all possible angles:

$$\pi F(\tau_T, \nu) = 2\pi \int_0^{\pi/2} I(\theta, \nu) \cos\theta \sin\theta d\theta$$

$$= \frac{\pi j_\nu}{2\kappa_\nu \tau_T^2} \left(2\tau_T^2 - 1 + (2\tau_T + 1)\exp[-2\tau_T]\right) \qquad (4.15)$$

Comparing (4.15) with the flux which emerges when $\tau_T = 0$, for which $\pi F(0, \nu) = (4\pi/3)j_\nu R$, we can estimate the escape probability, $P(\tau_T)$ as:

$$P(\tau_T) = \frac{\pi F(\tau_T, \nu)}{\pi F(0, \nu)} = \frac{3}{4\tau_T} \left(1 - \frac{1}{2\tau_T^2} + \left(\frac{1}{\tau_T} + \frac{1}{2\tau_T^2}\right)\exp[-2\tau_T]\right) \quad (4.16)$$

Note that as $\tau_T \to 0$; $P(\tau_T) \to 1$ and, as $\tau_T \to \infty$; $P(\tau_T) \to 3/4\tau_T$.

In a plane parallel nebula, of optical thickness τ, a reasonable approximation to the escape probability is:

$$P(\tau) = \frac{1}{\tau}\left(1 - \exp[-\tau]\right); \tau \leq e;$$

$$= \frac{1}{\tau(\ln\tau)^{1/2}}\left(1 - \exp[-\tau]\right); \tau > e; \qquad (4.17)$$

In this case therefore: as $\tau \to 0$; $P(\tau) \to 1$ and, as $\tau_T \to \infty$; $P(\tau_T) \to 1/\tau(\ln\tau)^{1/2}$.
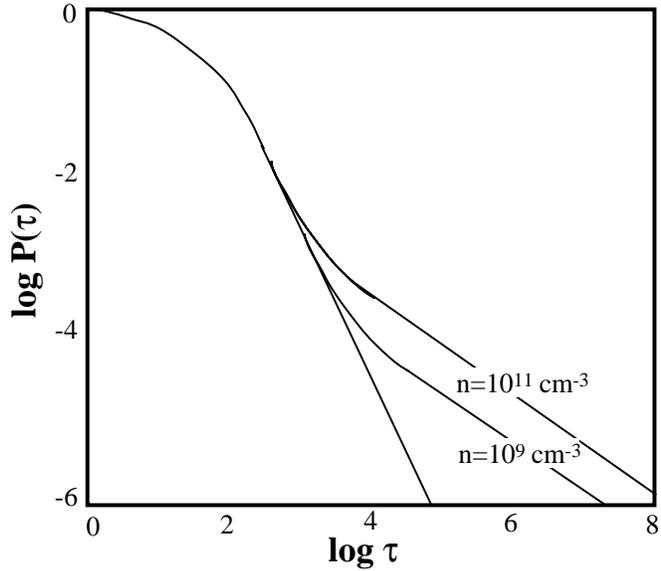
Fig. 4.5. The escape probability for the Lyα line as a function of the optical depth in the line and of the electron density. After Drake & Ulrich (1980).

In both these cases, and as also proves to be the case in other geometries, the escape probability at high optical depth varies approximately inversely as the optical depth, rather than as $\exp[-\tau]$, as we might have näively expected. The reason for this is that the photon can only escape if it is produced within the optically thin zone at the boundary of the nebula or at the boundary of the line profile, and the probability of that varies as $1/\tau$.

In detailed line transfer computations, the escape probability from the line in enhanced at high densities by pressure broadening of the natural width of the line, as is shown in Figure (4.4). Such densities are appropriate for the broad line regions around active galactic nuclei.

4.1.4 Line Transfer in the Lyman Series

As Figure (4.4) makes clear, enormous optical depths can be achieved in the Lyα line. In H II regions which are ionised by hot stars, the Lyα line optical depths are of order 10^4 while in collisionally excited and ionised plasmas found in the accretion disks near active galactic nuclei, Lyα line optical depths are estimated to rise to several million. In these conditions, all the other lines in the Lyman series become optically thick as well.

Consider, as the simplest example, the fate of Lyβ photons in such a nebula. Although the Lyα line itself is condemned to be emitted and re–absorbed for as long as it is trapped in the emitting region, or as long as the
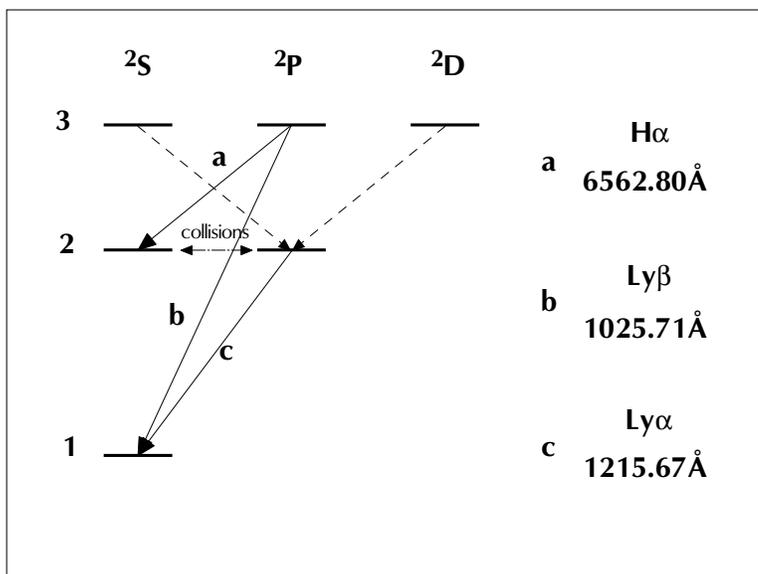
Fig. 4.6. Decay paths from the $3\,^2$P state of hydrogen. Lyβ photons will eventually be degraded into an Hα and a Lyα photon.

photon is not destroyed by some other means, the same is not true for the higher members of the Lyman series. However, it is clear from figure (4.5) that there exists another route for the radiative decay from the $n = 3$ level. For hydrogen, the relative probabilities of radiative decay from the excited $(n = 3)$ level are: $P(3^2\mathrm{P} - 1^2\mathrm{S}) = 0.882$; $P(3^2\mathrm{P} - 2^2\mathrm{S}) = 0.118$. Thus, the Ly$\beta$ photon has roughly a 12% chance of being destroyed in *each* scattering. After a few scatterings, the survival probability for a Lyβ photon becomes negligibly small, as Table 4.1.4 shows.

Table 4.2. Lyβ photon scattering and survival.

No. of Scatterings	1	5	10	100	1000
Survival Probability	0.882	0.534	0.285	3.5×10^{-6}	2.9×10^{-55}

All similar cascades resulting from the fluorescent conversion of higher members of the Lyman series must eventually produce a Lyα photon, since even were they to produce another Lyman series photon with lower n, this would quickly be converted in its turn. Thus, in nebulae, each photon belonging to a higher member of the Lyman series is rapidly converted to a Lyα photon plus members of other series (Balmer, Paschen, Pfund *etc.*).
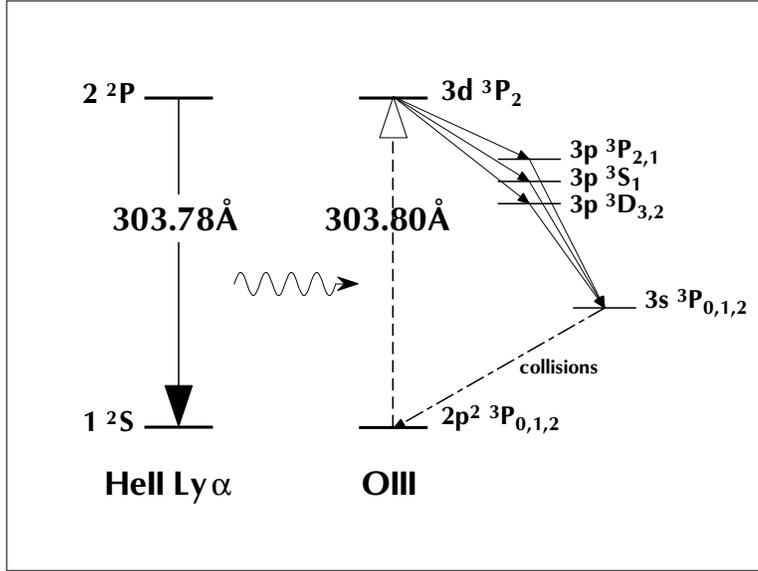
Fig. 4.7. The Bowen fluorescent process. This is active in highly–ionised regions such as in the inner regions of planetary nebulae, novae, or active galactic nuclei.

4.2 Fluorescent Processes

Fluorescence occurs where there is a coincidence (or near–coincidence) in wavelength between two transitions in different ions, in which case photons emitted by one ion can be re–absorbed by another and degraded into two or more photons of lower energy. This occurs in a number of nebular situations, and through a variety of physical processes which we will now briefly review.

4.2.1 The Bowen Mechanism

The near coincidence of the He II Lyα line at $\lambda 303.78$Å and the $2p^2$ $^3P_2$ – 3d $^3P_2$ resonance transition in [O III] at $\lambda 303.80$Å, allows conversion of some of the He II Lyα photons to O III photons in the 2800-3810Å region, notably the transitions at $\lambda\lambda 3133$, 3312, 3341, 3429, and 3444Å. This fluorescent cascade is shown in Figure (4.7). Since the number of fluorescent photons is related to the number of He II Lyα photons in the nebula, the Bowen mechanism has been used to estimate the rate of He ionisations. This places observational constraints on the shape of the EUV spectrum of the exciting object (planetary nebula nucleus, nova core, or the active galactic nucleus).

4.2.2 O I Fluorescence with Lyβ.

A close coincidence exists between the UV4 line of O I ($2p^4$ ^{3}P – 3d ^{3}D) at $\lambda 1025$Å and the Lyβ line at $\lambda 1026$Å. However, as we described above, the Lyβ line is itself fluorescently degraded after only a few scatterings. Therefore, in order to induce fluorescence in O I we require that neutral plasma is located immediately adjacent to ionised plasma producing a strong Lyβ line. Such a geometry occurs in the early evolution of nova shells, when the expanding ejected shell is just starting to become photoionised by the hot central core, or in young planetary nebulae with dense neutral inclusions. The optical signature of this fluorescence is the very strong $\lambda 8446$Å (3p ^{3}P – 3s ^{3}S) line produced as part of the fluorescent cascade back down to the ground state (3d ^{3}D $\stackrel{11287Å}{\rightarrow}$ 3p ^{3}P $\stackrel{8446Å}{\rightarrow}$ 3s ^{3}S$\stackrel{1302Å}{\rightarrow}$ $2p^4$ ^{3}P). This line, which is very prominent in the early spectra of nova shells fades dramatically as the shell expands and becomes more fully ionised.

4.2.3 H$_2$ Fluorescence with Lyα

Advanced Section

In the warm molecular gas surrounding a source of Lyα photons, a number of low lying excited H$_2$ transitions, notably the (1-2)P(5) line, can resonantly absorb Lyα and convert this to a variety of emission lines in the 1100Å– 1700Å region. As the simplified level diagram in Figure(4.7) shows, the process only works if the molecular hydrogen is in a vibrationally and rotationally excited state in the ground electronic state of molecular hydrogen. This requires the molecular hydrogen region to be heated to perhaps a couple of thousand degrees. This most likely requires that it has been recently shocked. However, the shock cannot have too high a velocity, since this would have dissociated the molecular hydrogen. At the same time, the molecular gas has to be in close contact with ionised plasma producing the Lyα line.

This process was first observed by Jordan *et al.* (1977) in spectra of sunspots, and has since been observed by Schwartz (1983) and Curiel *et al.* (1995) in low excitation Herbig–Haro objects. These objects are dense shocked regions around young stellar objects with fast ionised outflows, and so fulfill the rather stringent conditions for this process to work.

4.2.4 Raman Scattering Fluorescence

Advanced Section

When there is a near coincidence of an excited state of an abundant ion with a large oscillator strength, the cross–section for scattering is much enhanced. Although the atom, following absorption of the photon, is in a virtual state it nonetheless has an electron configuration which is very similar to that of the nearby bound state, and as a consequence, its lifetime is
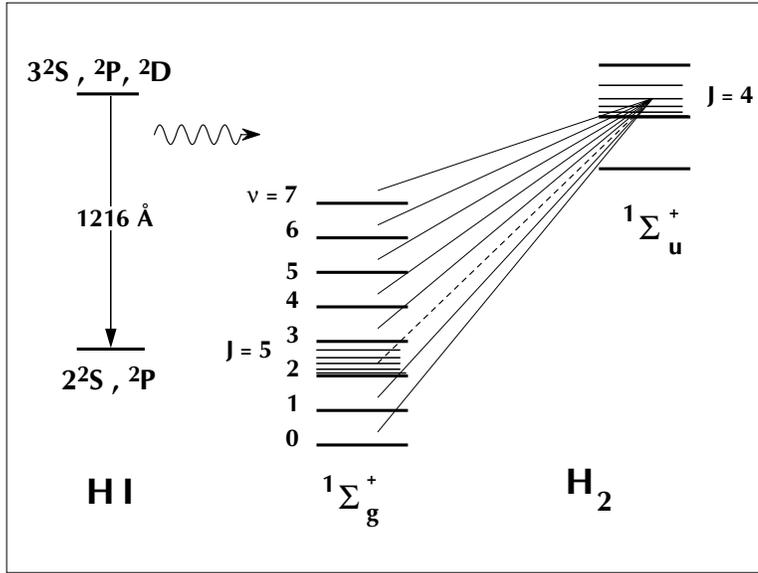
Fig. 4.8. Fluoresence of H_2 with Lyα. In order for this process to occur, the molecular hydrogen has to be both vibrationally and rotationally excited,which requires a temperature above about 1000 K
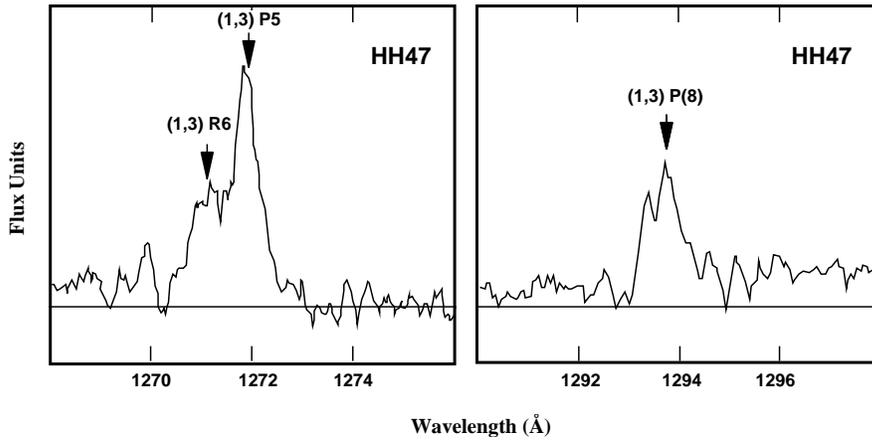
Fig. 4.9. Some fluorescent lines of molecular hydrogen observed in the Herbig–Haro object HH47 using the GHRS instrument on the Hubble Space Telescope (after Curiel *et al.* 1995)

enhanced, producing a rapid increase of scattering cross section as the virtual state approaches the bound state. This virtual state can then fluorescently decay by radiation to other excited states of the scattering atom or ion. The best example of this phenomenon is the Raman scattering of the O VI resonance lines by H I which is seen in Symbiotic stars (see 4.10). The theory was given by Schmid (1989).

If the incoming photon has frequency ν_i and the scattered photon as frequency ν_f then the cross–section for Raman Scattering, σ_R, is given in terms of the Thompson scattering cross–section, $\sigma_T = (8\pi/3)(\mathbf{e}^2/m_e c^2)^2 = 6.65 \times 10^{-25} \, \mathrm{cm}^2$, by:

$$\sigma_R = \sigma_T \frac{\nu_i}{\nu_f} \, |M|^2 \tag{4.18}$$

with:

$$|M| = \sum_m \frac{\nu_i \nu_f}{4} \left(\frac{gf_{1m} gf_{fm}}{\nu_{1m} \nu_{fm}} \right)^{1/2} \frac{(\nu_{1m} + v_{fm})}{(\nu_{1m} - v_i)(\nu_{fm} + v_i)} \tag{4.19}$$

Since the Doppler width of the incoming line (in energy space) is unchanged in the scattering, but the scattered photon has much lower energy, the apparent Doppler line width of the outgoing photons at $\lambda 6825$Åand $\lambda 6825$Åare larger in the ratio:

$$(\Delta\lambda/\lambda)_f = (\Delta\lambda/\lambda)_i (\lambda_f/\lambda_i). \tag{4.20}$$

This ratio is about 6.7 in the case of the O VI resonance. This increase in apparent line width is clearly visible in Figure (4.11). Both the identification of these two lines, and their unusual width were a great mysteries when these were first observed in the spectra of symbiotic stars (Allen, 1980).

4.3 Astrophysical Masers

4.3.1 Theory

Recall equation (4.3), which gives the absorption cross–section as a function of frequency for a line broadened by Doppler motions:

$$\sigma(\nu) = \frac{\pi \mathbf{e}^2}{m_e c} f_{12} N_1 \left[1 - \frac{g_1 N_2}{g_2 N_1} \right] \left(\frac{M}{2\pi KT} \right)^{1/2} \exp\left[\frac{-Mc^2 (\nu - \nu_0)^2}{2kT\nu_0^2} \right] \tag{4.21}$$

Consider the term in square brackets, which represents the correction due to stimulated emission. When the level populations are in their Boltzmann ratio, this term equals $(1 - \exp[-h\nu/kT])$. For transitions giving rise to lines in the optical, the correction factor due to stimulated emission is negligible.
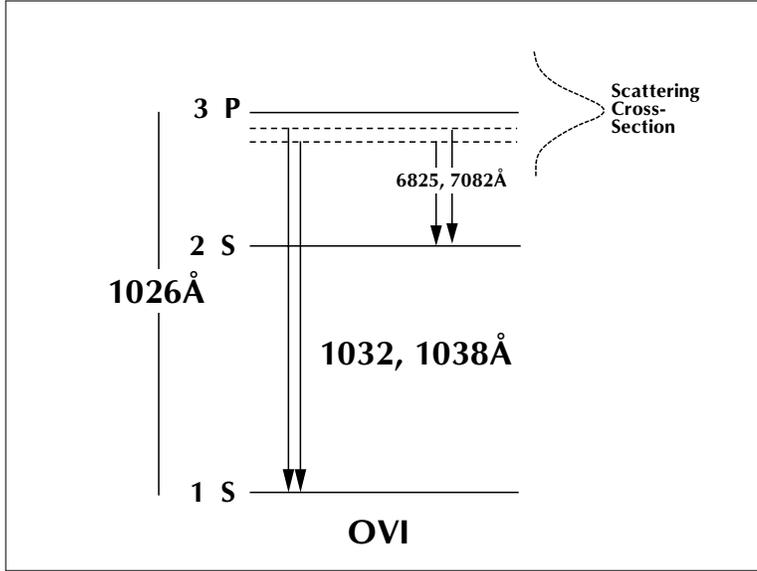
Fig. 4.10. Raman scattering of Lyα by O VI. This occurs in the interacting atmospheres of 'symbiotic' stars, which are close binary systems consisting of a hot white dwarf and a cool red giant star.

However, for lines in the microwave region, and at the temperatures commonly encountered in molecular clouds (10-100K), the correction factor may be quite large. Now, suppose that there is a third transition which is being pumped by collisions or by photons, such as is shown in Figure (4.12).

In general, transition rates vary as the cube of the frequency. Thus, if the pumping transition (1-3) and the cascade transition (3-2) are at a much higher frequency than the transition (2-1), then in most cases the transition rates in the pump (1-3) and (3-2) will be very much larger than the transition rate back to ground (2-1). This process therefore tends to build up the population in level (2) at the expense of level (1). If a *population inversion* $(g_1 N_2 > g_2 N_1)$ can occur through this process, then $1 - -g_1 N_2 > g_2 N_1 < 0$, which drives the effective absorption coefficient in equation (4.21) negative. That is to say, the light in the line is amplified rather than attenuated along the direction of the light propagation. Provided that the pumping rate is rapid, the light intensity increases exponentially over a distance determined by this negative absorption coefficient, called the maser gain G, *viz.* $I(x) = I_0 \exp[Gx]$. Such a maser is termed *unsaturated*. Interstellar masers therefore operate as *one–pass travelling–wave* masers without feedback. In such a cloud, the gain is a function of frequency:

$$G = G(\nu) = G_0 \exp\left[\frac{-Mc^2 \Delta\nu^2}{2kT\nu_0^2}\right] \tag{4.22}$$
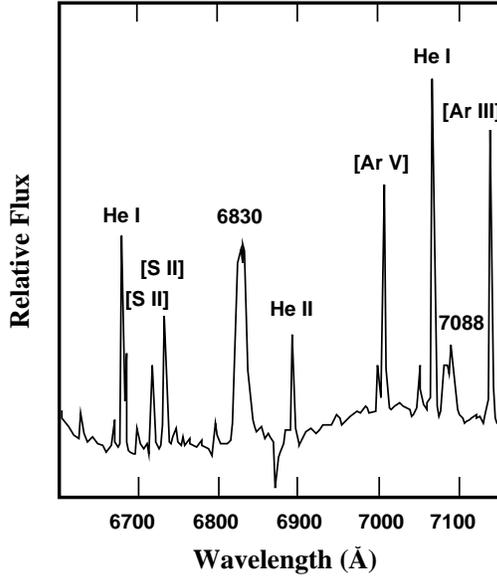
Fig. 4.11. The $\lambda 6830$ and $\lambda 7088$Å O IV fluorescent lines in the spectrum of V1016 Cygni, a symbiotic star. The large width of these lines compared with the other lines in the spectra is obvious. Note that the 7088Å line is weaker, because the virtual state which gives rise to it is further from the resonance line, and therefore has a lower Raman scattering cross section (after Schmid, 1989).

The consequence of this gain function is that the line is preferentially amplified closest to the line centre, so that the line width becomes progressively narrower than the Doppler width as it passes through the amplifying region. Figure (4.13) shows this effect as a function of the thermal e–folding line width, $(2kT\nu_0^2/Mc^2)^{1/2}$.

Equation (4.22) implicitly assumes that the pumping rate is sufficiently rapid that there is always a population inversion, regardless of how intense the maser has become. In practice, the pumping rate can become a limiting factor, and the amplification then only depends on the rate of pumping into the upper masing level. In this case, the amplification is determined only by the number of atoms processed through the masing beam of surface area $A(x)$. If this covers a solid angle Ω, then $G(x) \propto xA(x) \propto x^3\Omega$; since the area of the beam increases as x^2. Masers which are limited by the pumping rate are known as *saturated* masers. Such masers do not produce line narrowing, since the growth rate of the signal becomes uniform across the line.

Since the pumping rate is set by the rate for spontaneous emission in the pumping transition, saturation constrains the brightness temperature, T_B, of the maser to be less than:
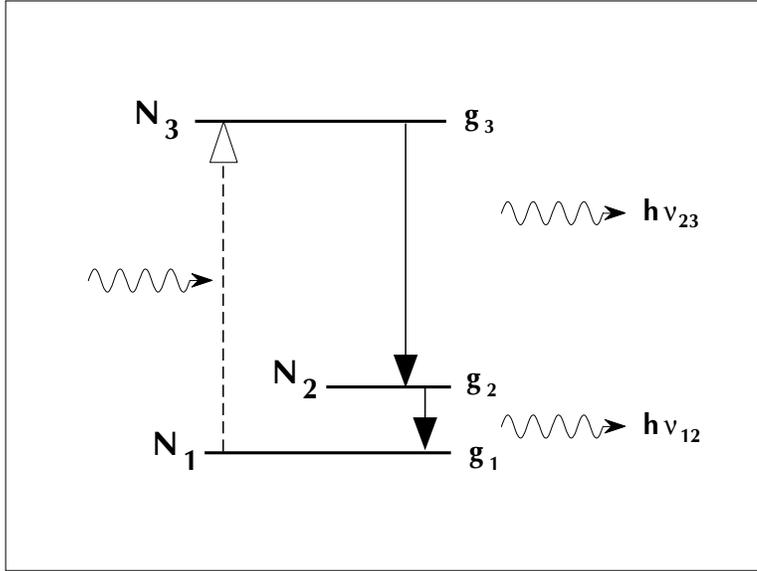
Fig. 4.12. Schematic of a masering transition. The transition between the two lower states is pumped by photon excitation into the upper state, followed by radiative decay to the second level. Under certain circumstances, this cycle is rapid enough to produce a population inversion between the two lower states.

$$T_B \leq \left(\frac{A_{13}}{A_{12}}\right)\left(\frac{4\pi}{\Omega}\right)\left(\frac{h\nu_{12}}{k}\right) \tag{4.23}$$

The first two factors determine the ratio of the brightness temperature to the thermal temperature. Thus, for a maser with $\Omega \sim 10^{-3}$ and $A_{13}/A_{12} \sim 10^7$, which might be considered 'typical' values, the brightness temperature in saturation could be as high as 10^{12} K! Strong interstellar masers are almost certainly saturated.

Since each maser photon requires at least one pump photon, we can relate the brightness temperature of the pump, as seen at the masing cloud to the brightness temperature of the masing source. This is called the Manley–Rowe relationship:

$$T_{pump} \geq \left(\frac{\Omega}{\Omega_{pump}}\right)\left(\frac{\nu_{12}}{\nu_{13}}\right) T_B \tag{4.24}$$

It is clear from this relationship that very high maser brightness temperatures can only be produced when the pumping source is close to the masing volume. Ideally the pump is generated by thermal processes, such as shocks, within the masing volume itself.
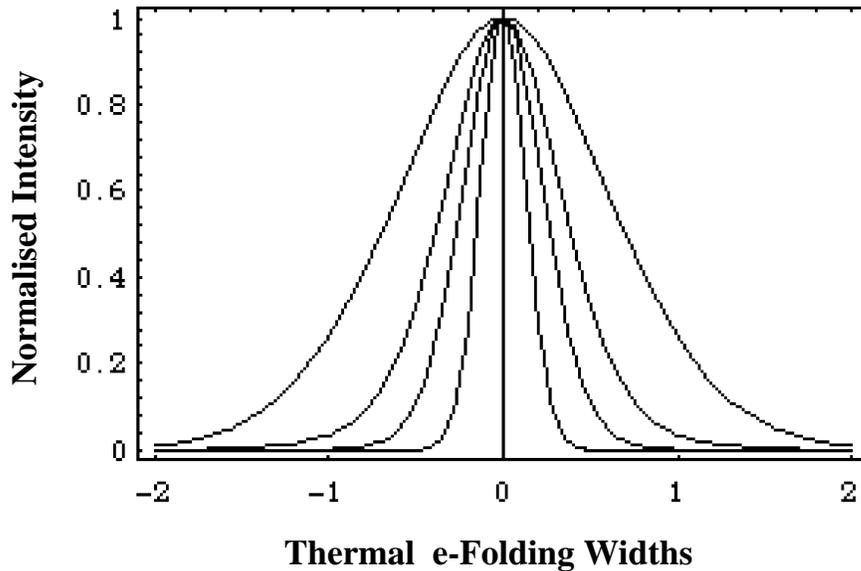
Fig. 4.13. Narrowing of the thermal line profile such as can occur for an unsaturated maser. Here the gain factors at the line centre, $G_0 x$, are given for 5, 10, 20 and 30.

4.3.2 Observations of Galactic Masers

Advanced Topic

Maser emission has been observed in eight different molecular species, OH, H_2O, SiO, HCN, CH, CH_3OH, H_2CO and NH_3, although not all of these are true interstellar masers. Here we will only discuss the OH and H_2O masers in detail. Of these, the OH sources show the highest brightness temperatures (as high as $T_B \sim 10^{15}$K!). Such brightness temperatures are only possible because they are produced in the outer atmospheres of mass–losing luminous old stars, where both the matter and radiation field densities are very high, and much higher than the interstellar medium in general. Maser sources are classified according to their *isotropic luminosity*, that is the luminosity that they would have if $\Omega = 4\pi$. Needless to say, this overestimates the true luminosity of the source by factors of possibly several thousand. Only a small fraction of the luminosity of the pumping source is converted to maser luminoisty. For example, a typical galactic OH source has an isotropic luminosity of only $10^{-3}L_{\odot}$, although the central star may be as luminous as $10^4 L_{\odot}$.

The mechanisms which drive the OH maser emission have been explained in detail by Elitzur (1992), *see* notes. Because the OH molecule is symmetric about the inter–nuclear axis, projections of the internal angular momenta on this axis (the z-axis) are conserved quantities. Since the projection of the end over end rotational angular momentum on this axis is always zero,

$J_z = L_z + S_z$. The electronic spin $S = 1/2$, and the ground electronic state is a Π-state ($L_z = 1$), so $J_z = 1 \pm 1/2$ which gives rise to two rotational ladders, $^2\Pi_{1/2}$ and $^2\Pi_{3/2}$. For each of these levels, the electronic interaction with the next electronic configuration, the Σ-state, removes the degeneracy between each parity state in the doublet, producing level splitting, the so–called $\Lambda-doubling$. The hyperfine interaction with the nuclear spin $\mathbf{I}$ further splits each member of the doublet into two further levels according to total angular momentum $\mathbf{F} = \mathbf{J} + \mathbf{I}$. The allowed transitions follow the dipole selection rules which require a parity change and $\Delta F = 0, \pm 1$ but with $F = 0 \nrightarrow 0$. The $\Delta F = 0$ lines at 1665 and 1667 MHz are called the *main* lines, while the $\Delta F = 1$ lines at 1612 and 1720 MHz are the *satellite* lines. Maser emission has been observed in all four levels as well as in the lines of some low–lying excited rotational states.

OH masers are found to be associated with the dusty mass–loss regions of evolved stars – the so–called OH/IR stars which are stars on the asymptotic giant branch stage of evolution, the last before the planetary nebular stage of evolution. The OH/IR stars are classified as *Type I* if they emit in the main lines, or as *Type II* if the 1612 MHz satellite lines are strongest. Type II sources generally emit in the main lines as well, but the 1720 MHz satellite line is never seen. This difference reflects the difference in the physical conditions and pumping in the maser excitation region in the two classes.

Because variations in the 1612 MHz emission in the Type II sources follow the variations in the radiation of the central star in lockstep, Harvey *et al.* (1974) were able to prove that this class of sources is pumped by rotational excitations caused by IR photons. This result is consistent with the absence of the 1720 MHz satellite line, which Elitzur (1976) showed can only be produced by collisions in a plasma with temperature less than about 200 K. This proportionality of maser luminosity and IR (pumping) luminosity also implies that the 1612 MHz OH masers are running in the saturated regime. The Type II masers show a "two–horn" velocity structure, consistent with the maser emission being produced in a relatively narrow, non–accelerating but expanding shell about the star. Shell radii are typically inferred to be in the range 10^{16} to 10^{17} cm.

The pumping mechanism of the main line Type I sources is more complex. To generate the required inversion of the ground–state $\Lambda-$ doublet requires preferential excitation of the upper $\Lambda-$ doublet components of the rotational ladders. This can be accomplished (Elitzur, 1978) by a radiation field whose photon ocupation number increases with frequency, and this can be generated by warm, optically thin dust emission. Detailed calculations show that dust temperatures in excess of 100 K are sufficient to produce main–line population inversion.

Observationally, the 1667 MHz emission line is stronger than, and occurs more often than, the 1665 MHz emission. This places limits on the dust temperature T_D in the range 150 K $< T_D <$ 280 K. The models imply that every

OH/IR star should exhibit main–line emission close to the central star, but that the Type II sources with the 1612 MHz line are produced only when the mass–loss rates from the central star are high.

OH maser activity is also seen in star–formation regions, in the vicinity of ultra–compact ionised hydrogen regions associated with newly–born massive stars. The masering regions have ages of only a few hundred thousand years. The individual spots of maser emission are only a few 10^{14} cm across, and their separations are similar, so it is probable that their observed size is their real physical size, although it is likely that they are filamentary along the line of sight to provide favourable conditions for amplification. The masering spots have individual velocity dispersions of order $1\,\mathrm{km\,s^{-1}}$, but the ensemble of spots seems to be formed in a dense expanding shell of compressed molecular gas surrounding the ionised region. This has an expansion velocity of a few $\mathrm{km\,s^{-1}}$. Extensive masering activity is also seen in the excited states of OH.

The details of the excitation mechanism in star fomring regions remains to be fully worked out. Unlike the case in OH/IR stars, conditions are not so suitable for radiative pumping – both the intensity of the radiation field and the dust temperatures are lower, but the maser luminosities are higher. This would be impossible if the maser is saturated as is the case in the OH/IR stars. It is likely therefore that collisional excitation is playing a key role in the masers seen in ultra–compact H II regions.

Masering activity is seen in the H_2O molecule in the OH/IR stars and is also associated with star formation in the vicinity of the ultra–compact H II regions. The pumping process which excites the H_2O masers is definitely collisional, and it requires excitation temperatures of order 1000 K.

The H_2O molecule is planar with an axis of symmetry passing through the O atom and between the two H atoms. It is clear that the moment of inertia in this axis and in the two orthogonal axes are all different. In rotation the molecule acts neither as a symmetric prolate rotator not as a symmetric oblate rotator – it is intermediate between these two limits. However, the rotational states are quantised according to the total angular momentum quantum number (J) and its projection on these two axes (K_- and K_+), and labelled as $J_{K_- K+}$. For example, in order of increasing energy, the lowest rotational level is split in two levels, 1_{01} and 1_{10}, the $J = 2$ state is split into two levels, 2_{12} and 2_{21}, the $J = 3$ state is split into four levels, 3_{03}, 3_{12}, 3_{21}, and 3_{30}, in order of increasing excitation energy and the $J = 4$ state is likewise split into four levels, 4_{14}, 4_{23}, 4_{32}, and 4_{41}.

The radiative selection rules require that K_- and K_+ must change their parity, and that $\Delta J = 0, \pm 1$. Thus, in collisional excitation of the molecule, within each J multiplet, radiative cascade down to the lowest–lying state is permitted, and this tends to build up a larger than LTE population in the $1_{10}, 2_{12}, 3_{03}, 4_{14}$ and higher states. This creates a population inversion between these states, and the radiatively accessible state of next lower J, leading to maser activity between these levels. The radiative feeding of this

excited state also leads to inversion of its population with respect to the next lower radiatively accessible J state, giving rise to a second masering transition, and so on for all radiatively accessible states. For example, masering activity can occur in all of the radiative sequence $7_{07} \rightarrow 6_{25} \rightarrow 5_{32} \rightarrow 4_{41}$. Many of the masering transitions of H_2O occur in the millimetric or sub–millimetric region of the spectrum, and are usually unobservable from earth, ironically thanks to absorption caused by the H_2O molecule in the earth's atmosphere. The first masering transition to be observed, the 22 GHz line of the $6_{16} \rightarrow 5_{23}$ transition just happens to occur at an unusually low frequency.

All late–type stars which display OH masering activity are thought to also contain H_2O masers, observable provided that sufficient sensitivity is used. The H_2O emission region is typically 15–50 times smaller than the OH emission region, and is subject to much greater (and more irregular) amplitude variability. This puts the masering region in the zone where the mass–loss wind is still being accelerated by radiation pressure acting on dust grains. The excitation is by neutral collisions (Deguchi, 1977; Cooke & Elitzur, 1985; Palma *et al.* 1988). The intensity of the maser luminosity is dependent on the mass–loss rate, and the inner boundary of the masering region is set by the point at which collisions thermalise the level populations.

The H_2O masers which are associated with newly–formed stars occur where there is a high–velocity outflow region interacting with the dense molecular cloud out of which the star has formed. These high velocity outflows are produced by the central star as it sheds the excess angular momentum of the accreted material (See Section (8.4.4)). Radial velocities and transverse motions of the masering spots (measured by very long baseline interferometry, VLBI), can easily range up to $100 \, \mathrm{km \, s^{-1}}$, and can be attributed to local turbulence and shearing flows within clouds in the size range $10^{14} - 10^{15}$ cm, and having lifetimes of only a few years. The brightness temperatures of these sources are enormous, typically ranging from 10^{11} to 10^{13} K. The sites of formation of the H_2O masers are most likely to be found in the dense cooling regions of shocks. These are either in the tails of fast, dissociative shocks where the chemistry favours the production of a large column density of warm ($\sim$ 400 K) water molecules (Elitzur, Hollenbach & McKee, 1989) and where the masers can be efficiently pumped (Neufeld & Melnick, 1991). Alternatively the sites are in slower, non–dissociative shocks (described in Section (8.5)) where the kinetic temperature is higher, in excess of 1000K, allowing for the pumping of highly excited levels. These two possibilities can be distinguished by observations in the sub–millimeter range, where higher transitions can be accessed. Such observations favour the idea that non–dissociative shocks are involved, since intensity ratios of masering transitions suggest a high kinetic temperature (Melnick *et al.* 1993).

4.3.3 Observations of Extragalactic Masers

Advanced Section

In extragalactic sources, masing activity has been observed in the OH, H_2O, H_2CO and CH molecules. Generally speaking the extragalactic sources are much more intense than the galactic ones, so a specialised "hyperspeak" has been developed to describe them. Extragalactic OH masers with isotropic luminosities in the range $1 - 1000\ L_\odot$ (10^3-10^6 times more luminous than galactic sources) are (somewhat inaccurately) called *megamasers*, and sources even more luminous than this are termed *gigamasers*.

The OH megamasers galaxies form a well defined subclass. The host galaxies are either luminous IR galaxies which are interacting or merging and therefore display very rapid rates of shock–induced star formation, or else they are feeding gas into a luminous (Seyfert) active galactic nucleus (AGN). In both of these classes, the re–processing of the stellar or AGN radiation field by a high column density of hot dust dominates the far–IR spectrum. In the most luminous of such sources the total dust luminosity can exceed $10^{12} L_\odot$!

The accepted model for the production of the megamaser activity was developed by Baan & Haschick (1984) and Baan (1985). In this, the maser represents low–gain, unsaturated amplification of the background radio continuum (originating either from the nucleus or from other radio sources in the galaxy) by the foreground molecular material. Despite the high luminosity of the OH maser, the required amplification factors are only a few (at most). The pumping is supplied by the far–IR photons, similar to the case of the galactic OH/IR sources.

Clear evidence in support of this scheme is provided by the fact that the properties of the OH megamasers are very closely tied to the properties of the host galaxy. Masing occurs when the slope of the spectrum in the 25 to $60\mu m$wavelength range is particularly steep; $\log(S_{25\mu m}/S_{60\mu m}) < -0.55$. There is also a close correlation between the OH and the IR luminosity; $L_{OH} \propto L_{IR}^2$ (Baan, 1989). It follows that the probability of detecting OH megamasers increases as the luminosity of the parent galaxy. This quadratic behaviour of the OH luminosity with the IR luminosity can be explained if the masers are unsaturated with relatively small gain and amplify the background radio continuum. In this case, the gain depends linearly on the OH column density and therefore the luminosity depends on the product of the background radio luminosity and the gain. However, observationally, the radio continuum is seen to depend linearly on the IR luminosity (with remarkably little scatter), and so if the OH column density also scales as the luminosity of the source, we recover the $L_{OH} \propto L_{IR}^2$ relationship.

The extragalactic H_2O maser sources fall into two classes, those which display galaxy–wide distributed emission of the same nature as the maser sources found in the Galactic star–formation regions, and those that are associated with active galactic nuclei (AGN). These are more than two orders of magnitude brighter. For example the famous case of NGC 4258 displays an isotropic luminosity of $120\ L_\odot$ and the luminous Seyfert NGC 1068 has

a isotropic luminosity of 350 $L_\odot$. This compares with the galactic star formation region W49 – the most luminous in the Galaxy – which can reach an isotropic luminosity $\sim 1 L_\odot$, at best. At the time of writing, some sixteen of these sources are known (Braatz, Wilson & Henkel, 1994, 1996; Koekemoer *et al.* 1995).

The nature of the H_2O maser emission in the AGN sources was recently reviewed by Maloney (1997). In many of them the emission arises in a dense, shocked molecular ring embedded in the accretion disk around the central object, near to the point at which the H_2O molecules become photo–dissociated. This ring is in Keplerian rotation within about one parsec of the central engine. The maser emission is seen in edge–on systems in regions where the velocity shear along the line of sight is low. This condition is satisfied directly in front of the central engine, where the masering amplitude is also assisted by the radio emission produced of the central engine and its associated radio emission regions. Low velocity shear also occurs at the tangent point in the orbiting ring of material, where the column densities may be much larger. These masers amplify the weaker radio emsision in the disk of the Galaxy. Masers produced here are either red–shifted or blue–shifted (by several thousand $\mathrm{km\,s^{-1}}$) by the Keplerian orbital motion, $v_{orb} = \sqrt{GM/R}$, where R is the radius of the molecular ring, and M is the mass interior to it.

The size of the observed orbital motion can be used to infer that there exists a massive Black Hole in the centre of such galaxies. For example, Miyoshi *et al.* (1995) and Greenhill *et al.* (1995) find that the mass enclosed within 0.2 pc of the centre of NGC 4258 is $2.1 \times 10^7 \, \mathrm{M_\odot}$. The central object inthis galaxy must be a Black Hole because the matter density implied by these numbers is in excess of $10^9 \, \mathrm{M_\odot \, pc^{-3}}$, and a cluster this dense cannot possibly remain stable for a timescale comparable to the age of the galaxy. (Figure 4.14) The Seyfert NGC 1068 provides a similar case, with the Black Hole mass inferred to be $\sim 10^7 \, \mathrm{M_\odot}$ (Greenhill *et al.* 1996). For the AGN sources, not only can the mass of the Black Hole be estimated, but the distance to the galaxy can be estimated to a remarkable accuracy ($\sim 4\%$) from the proper motion of the individual spots of H_2O maser activity as they pass in front of the central continuum source. At the same time, these features march steadily in velocity through the central emission complex (up to $600 \, \mathrm{km\,s^{-1}}$) thanks to the variation in $v_{rad} = v_{orb} \sin\theta$, where θ is the (small) angle measured along the orbit from the line of sight.

Notes on Chapter 4

- An excellent introduction to absorption line techniques and to atomic physics in the interstellar medium in general is that by Lyman Spitzer, Jr. 1978, *Physical Processes in the Interstellar Medium*, (Wiley:New York), ISBN 0-471-02232-2, which we have made extensive use of here. This book has recently been republished in the Wiley Classics Library Series.
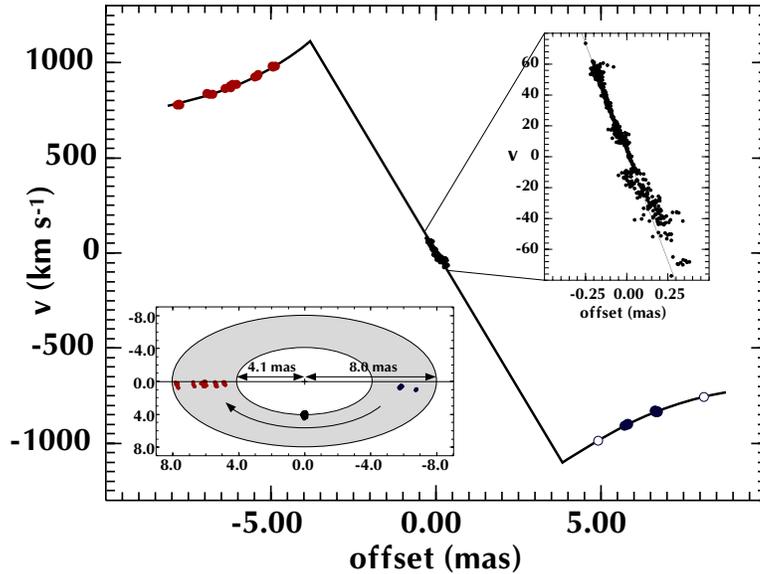
Fig. 4.14. NGC 4258 Rotation curve. (after Nakai et al 1995)

- An excellent recent review of interstellar line abundance analyses using the Hubble Space Telescope is by Savage, B.D., and Sembach, K.R. 1996, Ann. Rev. A&Ap, 34, 279.
- Recent compilations of oscillator strengths are to found in Verner, D, Barthel, P., and Tytler, D. 1994, ApJS, 62, 109 and Cardelli, J.A., Federman, S.R., Lambert, D.L., and Theodosiou, C.E. 1993, ApJL, 416, L41.
- A coherent account of the whole field of maser emission processes is that by Moshe Elitzur, 1992, *Astronomical Masers,* Astrophysics & Space Sciences Library v.170, Kluwer: Dordrecht, ISBN 0-7923-1216-3. This represents essential further reading for those interested in the field.

  A fairly recent review of extragalactic masers is by Henkel, C., Baan, W.A. & Mauersberger, R. 1991, Astronomy and Astrophysics Review, 3, 47. Progress in the whole field was also reviewed by Moshe Elitzur, 1992, Ann. Rev. A&Ap, 30, 75.

Exercises

Exercise 4.3.1. What happens to the local radiation intensity at a frequency corresponding to the centre of a strong, optically-thick resonance line produced by an excited state of an ion radiating to the ground state:

a. Within a nebula in which the upper level population is determined by collisions to and from the ground state and by the resonance line radiation itself (assume that the nebula is very optically-thick in the line).

b. Outside the nebula, where the radiation density in the resonance line is very low, but where the ion involved in the transition is the most abundant ion

c. Outside the nebula, where the radiation density in the resonance line is very low, but where the temperature is too low or too high to produce this ion in appreciable abundance

Exercise 4.3.2. A particular ion has an a line at 1548Å which has an oscillator strength of 3.0 (change to fit C IV?). Along a particular sight line, it is observed as an unsaturated absorption line with an equivalent width of 255mÅ.

a. What is the column density of this ion along the line of sight?

b. The line is broadened by turbulent motions to a velocity width of 60 $km.s^{-1}$. What then is the optical depth at the line centre?

Exercise 4.3.3. Assuming that the $J = 1 - 0$ transition in ^{12}CO at 115271.204 MHz is optically thick in a molecular cloud of radius 1.0pc, that the cloud has a temperature of 200 K, and an internal microturbulence of 1.3 $km.s^{-1}$, then what is the luminosity of the cloud in this CO line?

Exercise 4.3.4. An active galaxy is thought to have a massive Black Hole associated with it. An H_2O maser source associated with this nucleus shows three components, one at the (known) redshift of the centre of the galaxy, and the other two at ± 2200 $km.s^{-1}$ with respect this velocity.

a. Where do you think each of these components are formed?

The central component shows a complex structure, with subcomponents appearing at -50 $km.s^{-1}$ with respect to the central velocity, moving in velocity linearly with time, and eventually disappearing at $+50$ $km.s^{-1}$, about 1000 days after they first appeared.

b. How would you explain these observations?

c. In the light of your interpretation, what would you estimate the mass of the central black hole to be, in solar masses (1.998×10^{33} g)?

5. Collisional Ionisation Equilibrium

"Hangs in the uncertain balance of proud time"
— *Robert Greene*

Collisional ionisation equilibrium (sometimes called *coronal equilibrium* from its application to the hot coronae of stars) is a dynamic balance at a given temperature between *collisional ionisation* from the ground states of the various atoms and ions present in the plasma, and the processes of *recombination* from the higher ionisation stages. It is to be distinguished from the ionisation balance which is achieved under conditions of Local Thermodynamic Equilibrium (LTE). This equilibrium is described by the Saha Equation (frequently used in stellar theory). This describes a different equilibrium from the coronal approximation because in LTE the excited states are populated according to the Boltzmann law, which therefore play an important role in determining the ionisation state of the medium.

In practice, coronal equilibrium may never (or hardly ever) be achieved in real diffuse astrophysical plasmas, for reasons that are developed in the next chapter. However, its study offers many valuable physical insights into the ionisation balance of X–ray emitting plasmas, and in the past has been extensively (if often incorrectly) used in interpreting X–ray or UV line data obtained in astrophysical sources.

In this chapter we will develop the equations of collisional ionisation equilibrium, which is probably the simplest ionisation balance condition to understand. In order to do this, we must first review the various ionisation and recombination processes that we need to consider. Here we will emphasise the use of simple analytical fitting formulae for the various processes, since these provide the most convenient form for inclusion into computational codes. However, as accurate, self–consistent quantum mechanical calculations become increasingly available, more complex data will have to be incorperated into codes in numerical form, often as tabular data.

5.1 Collisional Ionisation

Direct *collisional ionisation* is the process whereby an electron strikes an ion, A (with charge $i+$), with sufficient energy to strip out a bound electron:
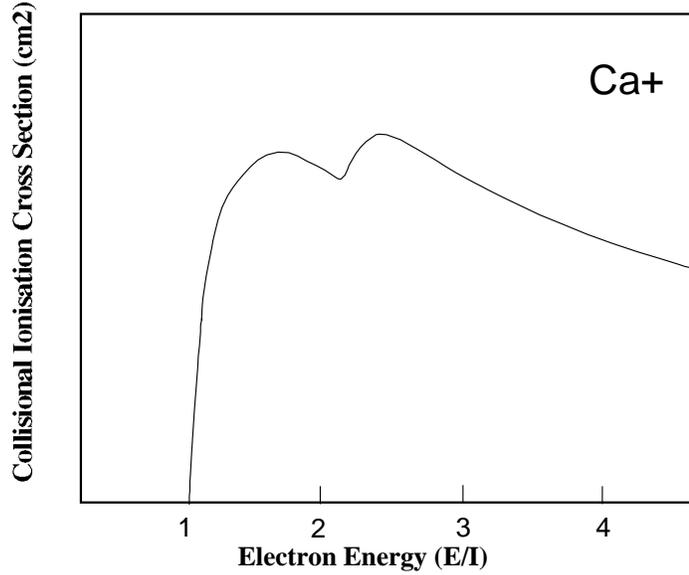
Fig. 5.1. The collisional ionisation cross–section for the Ca II, showing how the cross–section increases once more as the threshold for ionisation to an excited state is crossed.

$$A^{i+} + e^- \rightarrow A^{(i+1)+} + 2e^- - \Delta E_{A^{i+}} \qquad (5.1)$$
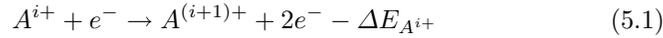
An energy equal to the ionisation potential of the atom, $\Delta E_{A^{i+}}$, is removed from the incoming electron. Collisional ionisation is therefore a process which effectively cools the electron gas. Note that the inverse process:

$$A^{(i+1)+} + 2e^- \rightarrow A^{i+} + e^- \qquad (5.2)$$
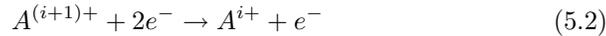
is known as *di–electronic recombination*. This process is discussed further below.

   With sufficient energy of impact, when the energy of the incoming electron exceeds the sum of the ionisation potential and the excitation potential, the ionisation may occur into an excited state of the more highly ionised ion, followed by a radiative return to the ground state. When the kinetic energy in the system exceeds the threshold energy for this process, a new channel for ionisation is opened up, and the cross–section for ionisation is correspondingly increased. This is illustrated in Figure (5.1) for the Ca II ion.

   The process of collisional ionisation can be considered as essentially similar to collisional excitation, except instead of excitation occurring to a single level, it occurs to a continuum of levels above the ionisation potential of the ion concerned. Provided that the electrons have a thermal Maxwell distribution, the collisional ionisation rate $R_{coll}^{A,i}$ ($\mathrm{cm^{-3}\,s^{-1}}$) for an atom A in its ith

ionisation stage is given in terms of the ionisation cross–section $\sigma_{coll}(E)$ and the number density of electrons, n_e, and ions, $n_{A,i}$, as (*c.f.* equation 3.3):

$$
\begin{aligned}
R_{coll}^{A,i} &= n_e n_{A,i} \alpha_{coll}^{A,i} \\
&= n_e n_{A,i} \int_I^\infty \sigma_{coll}(E).E.f(E)dE \; \mathrm{cm}^3 \, \mathrm{s}^{-1},
\end{aligned}
\tag{5.3}
$$

where $f(E)$ is the Maxwell distribution of the electrons in energy.

The cross–section is best calculated (Arnaud & Rothenflug 1985; Sutherland & Dopita 1993) using the methods of Younger (1981, 1982, 1983), based on a five parameter fit to each channel of the collision cross–section. These parameters can be derived either directly from collision cross–section experiments, or from theoretical calculations. The fits are made on a channel–by–channel basis which separates the contribution due to each electron configuration in the more highly ionised species. The cross–section for the mth. channel is expressed as (*c.f.* equation 3.8):

$$
u I_{A,i}^2 \sigma_{coll}(m,u) = A \left(1 - \frac{1}{u}\right) + B \left(1 - \frac{1}{u}\right)^2 + C \ln [u] + \frac{D}{u} \ln [u]
\tag{5.4}
$$

with $u = E/I_{A,i}$. Integrating the collisional ionisation cross–section over the Maxwell distribution at temperature T_e gives the *collisional ionisation rate*:

$$
\alpha_{coll}^{A,i}(n) = \frac{6.69 \times 10^{-7} x^{3/2}}{I_{A,i}^{3/2}} \int_1^\infty u I_{A,i}^2 \sigma_{coll}(m,u) \exp [-ux] \, du \; \mathrm{cm}^3 \, \mathrm{s}^{-1}.
\tag{5.5}
$$

with x_e being the ratio of the ionisation energy of the ion to the thermal energy of the electron gas, $I_{A,i}/kT_e$. The actual rate of collisional ionisations per unit volume is then given by the product

$$
R_{coll}^{A,i} = n_e n_{A,i} \alpha_{coll}^{A,i} \; \mathrm{cm}^{-3} \, \mathrm{s}^{-1}.
\tag{5.6}
$$

Using equation (5.4), equation (5.5) can be integrated explicitly over all the channels ($m = 1 \to m_{\max}$) which contribute to the total ionisation cross–section:

$$
\alpha_{coll}^{A,i} = \frac{6.69 \times 10^{-7}}{kT^{3/2}} \sum_{m=1}^{m=m_{\max}} \frac{\exp [-x_m]}{x_m} \Phi (x_m)
\tag{5.7}
$$

with :

$$
\begin{aligned}
\Phi (x) = {} & A + B \left(1 + x\right) - \left(C + Ax - B(2x - x^2)\right) E_1 (x) \exp [x] \\
& + D E_2(x) \exp [x]
\end{aligned}
$$

where E_1 and E_2 are the first and second Exponential integrals. For impact energies close to the threshold, $kT_e \ll I_{A,i}$, the contributions of higher energy channels ($n > 1$) can be neglected.

A major advantage of this method of solution for the collisional ionisation rate is that the coefficients show a regular progression along isoelectronic sequences. Thus, the coefficients may be estimated with good accuracy in those cases where laboratory of theoretical estimates might be lacking.

At sufficiently high electron impact energies, more than one electron of the target nucleus may be excited, leaving the atom in an unstable state, which is stabilised by the radiationless ejection of an electron, possibly followed by a radiative decay of the ionised atom back to its ground state:

$$A^{i+} + e^- \;\; \rightarrow A_*^{i+} + e^- - E_1$$

followed by :

$$A_*^{i+} \qquad \rightarrow A_*^{(i+1)+} + e^- + E_2 : A_*^{(i+1)+} \rightarrow A^{(i+1)+} + h\nu$$
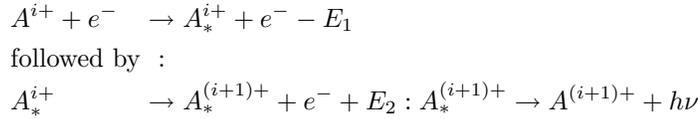
This process is known as *excitation–autoionisation*, and is favoured in heavy elements which have a large number of inner shell electrons and only one or two electrons in the outer shell. The complexity of this process means that is is often more difficult to give simple fitting formulae. However, for many ions excitation–autoionisation cross–sections of the following form provide an adequate fit:

$$\sigma_{EA}(u) = \frac{a}{u}\left(1 - \frac{1}{u^3}\right). \tag{5.8}$$

Here $u = E/I_{EA}$ and the effective excitation–autoionisation potential along a given iso–electronic sequence is given by $I_{EA} = E_0(Z - N)^p$ where Z is the atomic number, N is the number of inner–shell electrons in the isoelectronic sequence, and p is some power.

5.1.1 The Case of Hydrogen

The collisional cross–section for hydrogen according to equation (5.4) is shown in Figure 5.2. In this case there is, by definition, only one channel for ionisation. In fact, a reasonable approximation to this cross section can be obtained using the older semi–empirical Lotz formula, which was extensively used before good–quality theoretical and experimental cross sections were available. The Lotz formula is obtained by setting A, B and D in equation (5.4) equal to zero, and adopting $C = 4\times10^{-14}\,\mathrm{cm}^2\,\mathrm{eV}^2$. An even simpler (but still useful) approximation is obtained by a linear fit to the collisional ionisation cross–section in the region of the threshold: $\sigma_{coll,H} = \sigma_0(E - I_H)/I_H$ with $\sigma_0 \sim 10^{-16}\,\mathrm{cm}^2$ for hydrogen.

With this linear approximation, the collisional ionisation rate ($\mathrm{cm}^3\,\mathrm{s}^{-1}$) is then:

$$\alpha_{coll}^H(T_e) = 2.4849 \times 10^6 \frac{\sigma_0}{I_H} \int_{I_H}^{\infty} (E - I_H)E\exp\left[-I_H/kT_e\right] \tag{5.9}$$

where the numerical constant is $2^{7/2}(\pi m_e)^{-1/2}(kT_e)^{-3/2}$. After integration by parts, and substitution for σ_0 this gives:
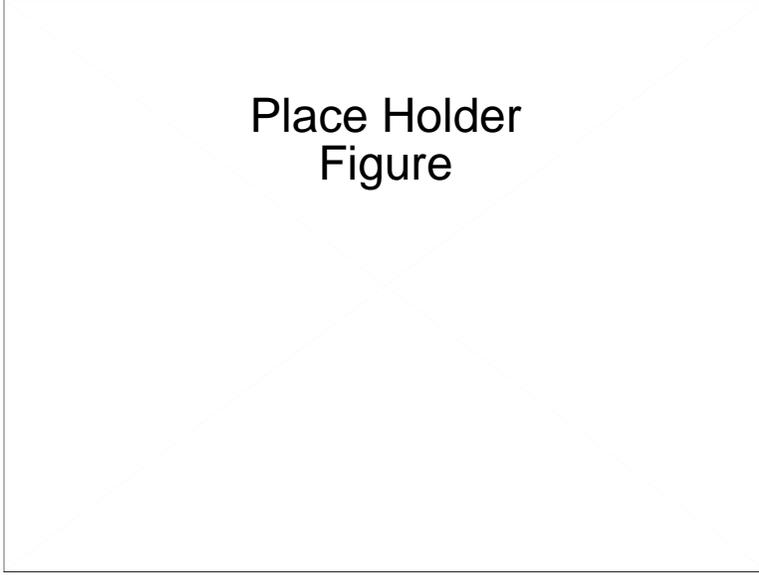
Fig. 5.2. The collisional ionisation cross–section for hydrogen calculated from equation (5.4) as a function of electron energy (measured in units of the ionisation potential, 13.6eV)

$$\alpha_{coll}^{H}(T_e) = 2.5 \times 10^{-10} \left(1 + \frac{T}{78,945}\right) T_e^{1/2} \exp\left[-157,890/T_e\right], \qquad (5.10)$$

and as before the actual ionisations occuring per second per unit volume is the product of of this rate and the electron and hydrogen densities

$$R_{coll}^{H} = n_e n_H \alpha_{coll}^{H}. \qquad (5.11)$$

The similarity of a collisional ionisation rate equation such as (5.10) and a collisional excitation rate equation such as (3.11) is obvious, the difference in the power–law dependence on temperature being simply related to the difference in the behaviour of the excitation and the ionisation cross–sections with energy above their respective thresholds.

[Table of collisional ionisation rate coeffs for some ions]

5.2 Recombination

Radiative recombination is the process of capture of an electron by an ion with the excess energy being radiated away in a photon. In most cases the electron is captured into an excited state, and usually into a state of large principal quantum number and high angular momentum state, so that the

recombined, excited, ion radiates several photons in a radiative cascade back down to the ground state:

$$A^{i+} + e^- \rightarrow A_*^{(i-1)+} + h\nu$$
$$A_*^{(i-1)+} \quad \rightarrow A^{(i-1)+} + h\nu_1 + h\nu_2 + h\nu_3 \ldots$$
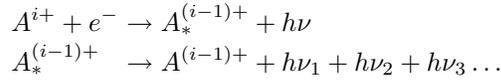
the photon in the first line represents a *recombination continuum* photon, since the energy of the impacting electron may take any value. However photons $\nu_1, \nu_2, \nu_3 \ldots$ represent quantised transitions between bound states, and are therefore termed *recombination lines*.

For any atom, as the principle quantum number, n, decreases, the number of states of different angular momentum quantum number, l, also decreases. However, the initial capture of the electron in recombination tends to occur into states of high angular momentum due to the initial orbital angular momentum of the electron around the ion. In the cascade to lower states, the selection rules of quantum mechanics force $\Delta l = \pm 1$. Electrons which find themselves in low angular momentum states can make either kind of transition. However, electrons which are initially in a high angular momentum state tend to get trapped into the highest l state consistent with the principal quantum number, and are thereby forced to cascade down to the ground state though transitions of the kind, $\Delta n = -1, \Delta l = -1$. Thus, in a recombining gas, within any $n-$ state of any ion, the l sub–levels are not populated in proportion to their statistical weights, as would be the case in a plasma in Local Thermodynamic Equilibrium (LTE). In addition, the n levels are not populated in anything like that expected from the Boltzmann Equation. To describe the actual level populations in the excited states, we generally use an *occupation factor* defined by the ratio of the population in the excited state compared with the population this state would have under LTE conditions at the same electron temperature and density; n_e, T_e, respectively:

$$b_{nl}(n_e, T_e) = N(n, l)/N_{\text{LTE}}(n, l) \tag{5.12}$$

The factor $N_{\text{LTE}}(n, l)$ can be readily derived. In LTE at the ionisation balance is given by the Saha Equation:

$$\frac{n_{A,i+1} n_e}{n_{A,i}} = \left(\frac{2\pi m_e k T_e}{h^2} \right)^{3/2} \exp\left[-\frac{I_{A^{i+}}}{kT_e} \right] \tag{5.13}$$

where $I_{A,i}$ is the ionisation potential of the ion A^{i+}. On the other hand, the population of the n, l sub–level with excitation energy $E_{A,i+1}(n, l)$, will be given by the Boltzmann equation:

$$\frac{n_{A,i+1}^{\text{LTE}}(n, l)}{n_{A,i}} = (2l + 1) \exp\left[-\frac{E_{A,i+1}(n, l)}{kT_e} \right] \tag{5.14}$$

so therefore it follows that:

$$n_{A,i+1}^{\text{LTE}}(n,l) = (2l+1)n_e n_{A,i+1} \left(\frac{2\pi m_e kT_e}{h^2} \right)^{-3/2} \exp\left[\frac{x_n}{kT} \right] \qquad (5.15)$$

with $x_n = I_{A,i} - E_{A,i}(n,l) \sim I_{A,i}/n^2$.

5.2.1 Radiative Recombination Rates

For a given ion of an element, A^{i+}, the radiative recombination rate to level n per unit time and volume is given in terms of the number densities of the more highly ionised ion $n_{A,i+1}$ and of the electrons, n_e, and in terms of the electron capture cross–section to level n of the ion A^i at a given electron velocity v; $\sigma_{rec}^{A,i}(n,v)$:

$$Rrec^{A^i}(n,T_e) = n_e n_{A,i+1} \int_0^\infty \sigma_{rec}^{A,i}(n,v) \left[1 + \frac{c^3 \rho(\nu)}{8\pi h\nu^3} \right] vf(v)dv \, \text{cm}^3 \, \text{s}^{-1}$$

$$= n_e n_{A,i+1} \alpha_{rad}^{A,i}(n,T_e) \qquad (5.16)$$

where $\rho(\nu)$ is the radiation density at the frequency ν, of the recombination continuum photon produced by recombination to level n, and $f(v)$ is the Boltzmann distribution function for the electrons at electron temperature T_e. In this equation, the second term in square brackets represents the *radiation–induced recombination*, analagous to the stimulated emission term in emission line theory. In practice, this term is not usually important in diffuse astrophysical plasmas, and is usually neglected.

The total effective recombination rate to all states can then be written as the sum of the recombination rate to each state:

$$\alpha_{rad}^{A,i}(T_e) = \sum_n \alpha(n,T_e) \qquad (5.17)$$

For hydrogenic ions, the radiative recombination rates can be calculated using a formula due to Seaton (1959) based on a expansion of the Kramers–Gaunt factor:

$$\alpha_{rad}^{A^i}(T_e) = 5.197 \times 10^{-14} Z\lambda^{1/2} \left(0.4288 + 0.4\ln[\lambda] + 0.469\lambda^{-1/3} \right) \, \text{cm}^3 \, \text{s}^{-1}.$$
$$(5.18)$$

where $\lambda = I(A^i)/kT_e = 157890Z^2/T_e$. For hydrogen, this can be approximated by a simple power law;

$$\alpha_{rad}^H(T_e) = 4.18 \times 10^{-13} [T/10^4 \, \text{K}]^{-0.72} \, \text{cm}^3 \, \text{s}^{-1}, \qquad (5.19)$$

which is applicable in the range 5000–20000 K.

For non–hydrogenic atoms, the usual way to compute radiative recombination rate is to evaluate the recombination rate to the ground state using the Milne relation (see Section 5.3.3), and then to compute the recombination

rate to the excited states using the hydrogenic approximation. The result is then often fitted to a simple power law with temperature:

$$\alpha_{rad}^{A^i}(T) = A_{rad}[T/10^4 K]^{-\eta} \, \text{cm}^3 \, \text{s}^{-1}. \tag{5.20}$$

[Table of recombination rate coeffs for some ions]

5.2.2 Di–electronic Recombination Rates

As mentioned above, *di–electronic recombination* is the inverse of *collisional ionisation*. The main pathway for di–electronic recombination is through an excitation of a core electron with capture of the passing electron;

$$A^{i+}(1s,\ldots) + e^- \rightarrow A_*^{(i-1)+}(n_1l_1; n_2l_2) \tag{5.21}$$
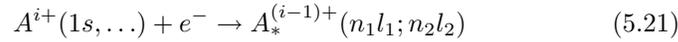
Here, one of the electrons is in an autoionising state, n_1l_1, and the other is in an excited state, n_2l_2. The ion first stabilises itself by one of the valence electrons radiating back to the ground state:

$$A_*^{(i-1)+}(n_1l_1; n_2l_2) \rightarrow A_*^{(i-1)+}(n_3l_3; n_2l_2) + h\nu \tag{5.22}$$
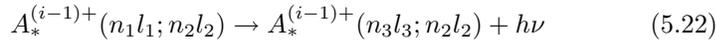
and then the ion is free to return to its ground state by radiative cascade:

$$A_*^{(i-1)+}(n_3l_3; n_2l_2) \rightarrow A^{(i-1)+}(n_3l_3; n_4l_4) + h\nu_1 + h\nu_2 + \ldots.. \tag{5.23}$$
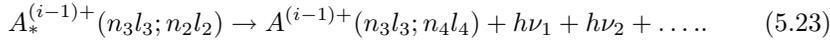
At high temperatures, core relaxation is the most important stabilising process, while at low temperatures the electron is captured *via* low–lying resonance states. There are thus two contributions to the total di–electronic recombination rate:

$$\alpha_{\text{diel.}}^{A,i}(T_e) = \alpha_{Lo-T}^{A,i}(T_e) + \alpha_{Hi-T}^{A,i}(T_e) \, \text{cm}^3 \, \text{s}^{-1}. \tag{5.24}$$

A fitting formula for low temperatures is given by Nussbaumer & Storey (1983):

$$\alpha_{Lo-T}^{A,i}(t) = 10^{-12} t^{-3/2} \left(\frac{a}{t} + b + ct + dt^2 \right) \exp\left[-\frac{f}{t} \right] \, \text{cm}^3 \, \text{s}^{-1}, \tag{5.25}$$

with $t = T_e/10^4$ K; a, b, c, d and f being fitting constants for each ion. At high temperatures the Shull & Van Steenberg (1982) form can be used:

$$\alpha_{Hi-T}^{A,i}(t) = A_{diel} t^{-3/2} \left(1 + B_{diel} \exp\left[\frac{-t_0}{t} \right] \right) \exp\left[-\frac{t_1}{t} \right] \, \text{cm}^3 \, \text{s}^{-1}, \tag{5.26}$$

with A_{diel}, A_{diel}, t_0 and t_1 being fitting constants for each ion.

In general, the low–temperature di–electronic recombination contribution is important at temperatures of about 1000–3000 K and the high temperature term is dominant above 20000 K. Single electron recombination dominates at very low temperatures and is usually an important contributor in the region 5000–20000 K. This is illustrated in Figure (5.3) for the case of carbon ions, taken from the fully self–consistent quantum mechanical calculations of Nahar & Pradhan (1997).

[Table of dielectronic rate coeffs for some ions]
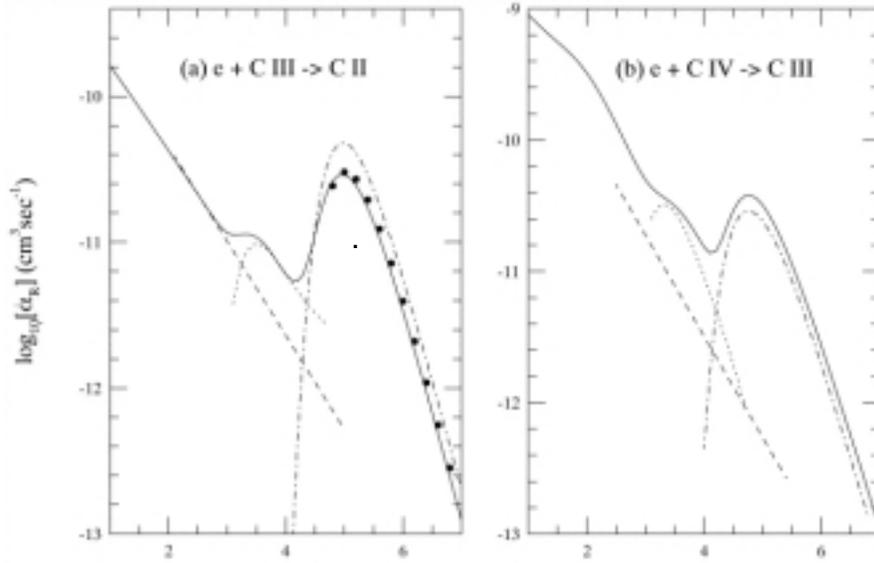
Fig. 5.3. Total recombination rate coefficients for C II, C III, C IV and C V. The dashed line is the radiative recombination from a fit of the form (5.14), the dotted line represents the low temperature di–electronic term (5.17), and the dot–dashed line is the high temperature part fitted by (5.18). The fitted circles are another calculation of this same contribution (after Nahar & Pradhan, 1997).

5.3 Photoionisation

5.3.1 From Outer Shells

Photoionisation is the inverse process to radiative recombination. As a consequence, there is an intimate connection between the radiative recombination cross–sections and the photoionisation cross section. This will be discussed later. Photoionisation, as the name suggests, is the ionisation of an atomic species by the absorption of a photon:

$$A^i + h\nu \rightarrow A^{(i+1)+} + e^- + \Delta E \qquad (5.27)$$
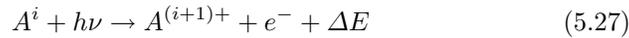
If the incoming photon has sufficient energy, it may leave the more highly ionised species in an excited state which subsequently decays by a radiative cascade back to down to the ground state:

$$A^i + h\nu \rightarrow A_*^{(i+1)} + e^- + \Delta E$$
$$A_*^{(i+1)} \rightarrow A_*^{(i+1)} + h\nu_1 + h\nu_2 \ldots$$
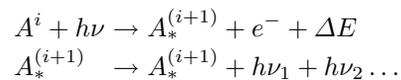
In this way, an appreciable number of additional channels to the photoionisation process become energetically available at higher energy, thus increasing the photoionisation cross–section. For example:

$$O^0(2p^4 \ ^3P) + h\nu \rightarrow O^0(2p^3 \ ^4S) + e^- : h\nu > 13.6eV$$
$$O^0(2p^4 \ ^3P) + h\nu \rightarrow O^0(2p^3 \ ^2D) + e^- : h\nu > 16.9eV$$
$$O^0(2p^4 \ ^3P) + h\nu \rightarrow O^0(2p^3 \ ^2P) + e^- : h\nu > 18.6eV$$
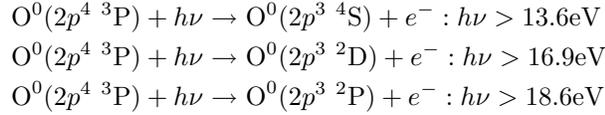
in this example, the ionisation occurs to different terms of the same electron configuration.

5.3.2 From Inner Shells

If we increase the energy of the incoming photon still further, it becomes possible to remove one of the inner shell electrons by *inner shell photoionisation* which also results in a change the electron configuration in the excited species, *i.e.* :

$$O^0(1s^22s^22p^4 \ ^3P) + h\nu \rightarrow O^+(1s2s^22p^4 \ ^2P \text{ or } ^4P) + e^- : h\nu > 544\,eV \quad (5.28)$$
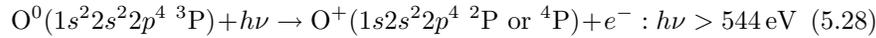
This may be followed by a radiative re–adjustment back to the ground state. However, in this particular case another mode of photoionisation becomes not only energetically possible, but indeed more probable; that of *Auger ionisation*. This is a photoionisation from an inner K– or L–shell, followed by a radiationless *autoionisation*, and is completed by radiative cascade back down to the ground state:

$$
\begin{aligned}
A^i + h\nu \ &\rightarrow A_{**}^{(i+1)+} + e^- + \Delta E_1 \\
A_{**}^{(i+1)} \ &\rightarrow A_*^{(i+m+1)} + me^- + \Delta E_2 \\
A_*^{(i+m+1)} \ &\rightarrow A^{(i+m+1)} + h\nu_1 + h\nu_2 \dots
\end{aligned}
$$
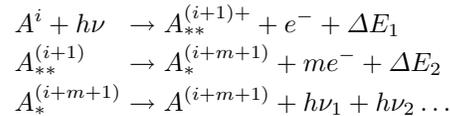
For example, in the case of the oxygen inner–shell photoionisation given above, the Auger ionisation path is (mostly) into the ^3P excited state of O III:

$$O^+(1s2s^22p^4 \ ^2P \text{ or} ^4P) \rightarrow O^{++}(1s^22s^22p^2 \ ^3P \ ) + e^- \quad (5.29)$$
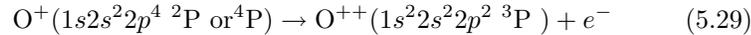
followed by radiative transitions to the ground state. In some cases these radiative transitions can affect the intensity of the weak lines which are normally used as temperature or density diagnostics in objects such as Active Galactic Nuclei. This was pointed out by Aldovandi & Gruenwald (1985). This process has recently been investigated for carbon, nitrogen and oxygen lines in a series of papers by Petrini and his co–workers. (Petrini & Da Silva 1997, Petrini & Farras 1994, Petrini and Da Aranjo 1994)

Note that the radiationless autoionisation may produce more than one electron. In general, m becomes greater than unity above some threshold energy, and increases as more channels for the Auger process become energetically accessible.

In inner shell photoionisation followed by Auger ionisation, two high energy electrons are produced, the first from the primary photoionisation, with

an energy equal to the photon energy less the binding energy of the ejected electron, and the second from the Auger ionisation. This has an energy which is roughly equal to the difference in binding energies of electrons in the inner shell and outer shells less the binding energy of the outer shell electron. This could be several hundred eV. Such fast, *supra–thermal* electrons are capable of either heating the gas, or of collisionally exciting and/or ionising ions with excitation or ionisation potential less than the energy of the fast electron. This secondary ionisation process becomes quite important when the supra–thermal electrons find themselves in a partially ionised zone. In this case, much of the energy of the fast electron is lost by collisional excitation and ionisation of hydrogen and, to a lesser extent, of neutral helium.

Analytic fits to Monte–Carlo models of this process have been given by Shull & Van Steenberg (1985) as a function of the fractional H– ionisation of the medium, x. The fraction of the primary photoelectron energy lost as heat is given by:

$$F_{Heat} = C \left[1 - (1 - x^a)^b \right] \tag{5.30}$$

with C, a and b equal to 0.9971, 0.2662 and 1.3163, respectively. For the fraction of primary photoelectron energy lost through the other processes, a formula of the form:

$$F = C \left(1 - x^a \right)^b \tag{5.31}$$

gave a good description of the results. The coefficients (C, a, b) are (0.3908, 0.4092, 1.7592) for collisional ionisation of hydrogen, (0.0554, 0.4614, 1.6660) for collisional ionisation of helium, (0.4766, 0.2735, 1.5221) for collisional excitation of hydrogen Lyα, and (0.0246, 0.4049, 1.6594) for collisional excitation of the He I Lyα line. The average number of collisional ionisations and excitations induced per fast electron can be easily calculated using the energy lost by each of these processes; 10.2eV and 13.6eV for collisional excitation and ionisation of hydrogen, respectively, and 19.95eV and 24.58eV for the corresponding processes in helium.

5.3.3 The Milne Relation

Clearly, the inverse process of photoionisation from a given level in an atom or ion is recombination by radiative capture of an electron into this same level. The intimate relation between the two processes emphasises the need to treat them both self–consistently and with the same set of atomic eigenfunctions. The Milne relation is a formula which relates the cross section for photoionisation with the cross–section for recombination, and was derived using the principle of *detailed balance* in the condition of Local Thermodynamic Equilibrium (LTE). In this condition, detailed balance requires that all the rates of the various atomic processes are balanced by the rates of their inverse processes. For a particular atom, ion and state with ionisation energy $I_{A,i} = h\nu_I$, the principle of detailed balance applied to photoionisation–recombination

states that the rate of spontaneous + radiatively induced recombinations of electrons in the velocity range $v \rightarrow v + dv$ is matched by the photoionisation rate in the frequency range $\nu \rightarrow \nu + d\nu$, where $m_e v^2/2 + h\nu_I = h\nu$. As a consequence of this, $m_e v dv = h d\nu$. In LTE, the radiation field density is given by the Black–Body value, and the induced radiative recombination rate is a factor $\exp[-h\nu/kT_e]$ of the photoionisation rate. The detailed balance equation between the number of atoms of atomic species A in their $(i+1)^{th}$. stage of ionisation; $n_{A,i+1}$, and the number of atoms of atomic species A in their i^{th}. stage of ionisation; $n_{A,i}$ is therefore:

$$n_e n_{A,i+1} \sigma_{rec}(v) f(v) dv = n_{A,i} \left(1 - \exp[-h\nu/kT_e]\right) \frac{4\pi B_\nu(T_e)}{h\nu} \sigma_{phot}(\nu) d\nu$$

(5.32)

where $\sigma_{phot}(\nu)$ is the photoionisation cross–section, $\sigma_{rec}(v)$ is the recombination cross–section for electron velocity v, $f(v)$ is the Boltzmann Distribution in velocity at temperature T_e :

$$f(v) = \frac{4}{\pi^{1/2}} \left(\frac{m_e}{2kT_e}\right)^{3/2} v^2 \exp\left[-\frac{m_e v^2}{2kT_e}\right]$$

(5.33)

and $B_\nu(T_e)$ is the Planck Function:

$$B_\nu(T_e) = \frac{2h\nu^3}{c^2} \left(\exp[-h\nu/kT_e] - 1\right)^{-1}$$

(5.34)

However, in LTE the Saha equation of ionisation balance also applies:

$$\frac{n_e n_{A,i+1}}{n_{A,i}} = \frac{2g_{i+1}}{g_i} \left(\frac{2\pi m_e kT_e}{h^2}\right)^{3/2} \exp\left[-\frac{h\nu}{kT_e}\right]$$

(5.35)

where here g is the statistical weight of a state. Combining all of these equations, we obtain the Milne Relation between the two cross–sections:

$$(m_e c v)^2 \sigma_{rec}(v) = \frac{g_i}{g_{i+1}} (h\nu)^2 \sigma_{phot}(\nu)$$

(5.36)

5.3.4 Photoionisation Cross–sections

Because of the intimate connections between the ionisation process, *collisional ionisation* and *photoionisation* and their inverse processes, *di–electronic recombination* and *radiative recombination*, respectively, the rate calculations of all of these processes should be computed in a fundamentally consistent manner using the same set of atomic eigenfunctions. Since such calculations of photoionisation cross–sections would automatically include autoionising resonances, the electron–ion recombination rates would also include both the radiative and the di–electronic recombination processes. In the past, these

have usually been treated separately, and by different methods, and (where possible) made self–consistent by applying the Milne relation.

The calculations of Nahar & Pradhan (1997) represent the first such fully–self consistent computations using the $R-$matrix method developed for the Opacity Project and the Iron Project (Berrington *et al.* 1987; Hummer *et al.* 1993) and extended by Nahar & Pradhan (1994). We have already shown the computations of the carbon ion recombination rates in Figure (5.3), above, and the corresponding photoionisation rates are shown in Figure (5.4).

When such calculations become available for more ions, then theoretical models of the ionisation state of the interstellar medium will have to include these in tabular form for each ion of each element. For the time being, however, a parametric fit in the same spirit as those already presented for the other processes can also be applied to photoionisation cross–sections.

For H–like ions, the photoionisation cross–section, $\sigma_{1s}(E)$, and the threshold energy, E_{1s}, can be calculated analytically:

$$E_{1s} = Z^2 I_H \text{ and}$$
$$\sigma_{1s}(\nu) = \frac{2^9 \pi^2 \alpha a_0^2}{3Z^2 \mathbf{e}^4} \left(\frac{E_{1s}}{h\nu}\right)^{3.5} = \frac{6.3042 \times 10^{-18}}{Z^2} \left(\frac{\nu}{\nu_0}\right)^{-3.5} \text{cm}^2, \quad (5.37)$$

where I_H is the ionisation potential of hydrogen, one Rydberg (13.6 eV), ν_0 is the associated frequency, α is the Fine Structure Constant, and a_0 is the Bohr Radius ($h^2/4\pi me^2 = 0.529 \times 10^{-8}$ cm).

For other atoms and ions, an interpolation formula of the form given by Daltabuit & Cox (1972) is frequently used:

$$\sigma_{phot}(\nu) = \sigma_0 \left[\beta \left(\frac{\nu}{\nu_0}\right)^{-s} + (1-\beta) \left(\frac{\nu}{\nu_0}\right)^{-(s+1)}\right] \quad (5.38)$$

where ν_0 is the frequency at threshold, and σ_0, β and s are fitting parameters.

For inner (K–shell) ionisation, Band *et al.* (1990) present a very useful set of fitting formulae using a modification of the hydrogen–like formula, based on the similarity of photoionisation cross–sections for a given shell, but across different atoms and ions. Here:

$$\sigma_{phot}(E) = \sigma_1 y^{-3/2} \left(1 + y^{1/2}\right)^{-4} : y = \frac{E}{E_1} \quad (5.39)$$

where the fitting parameters σ_1 (in mega–Barns; 10^{-18} cm^2) and E_1 are given in terms of the atomic number Z, and the number of bound electrons in the ion, N, by:

$$E_1(eV) = a_1 + a_2 Z + a_3 Z^2 + (a_4 + a_5 Z)N$$
$$\sigma_1(Mb) = \left(b_1 + b_2 Z + b_3 Z^2 + b_4 N\right)^{-1} \quad (5.40)$$
$$\text{with} \quad :$$
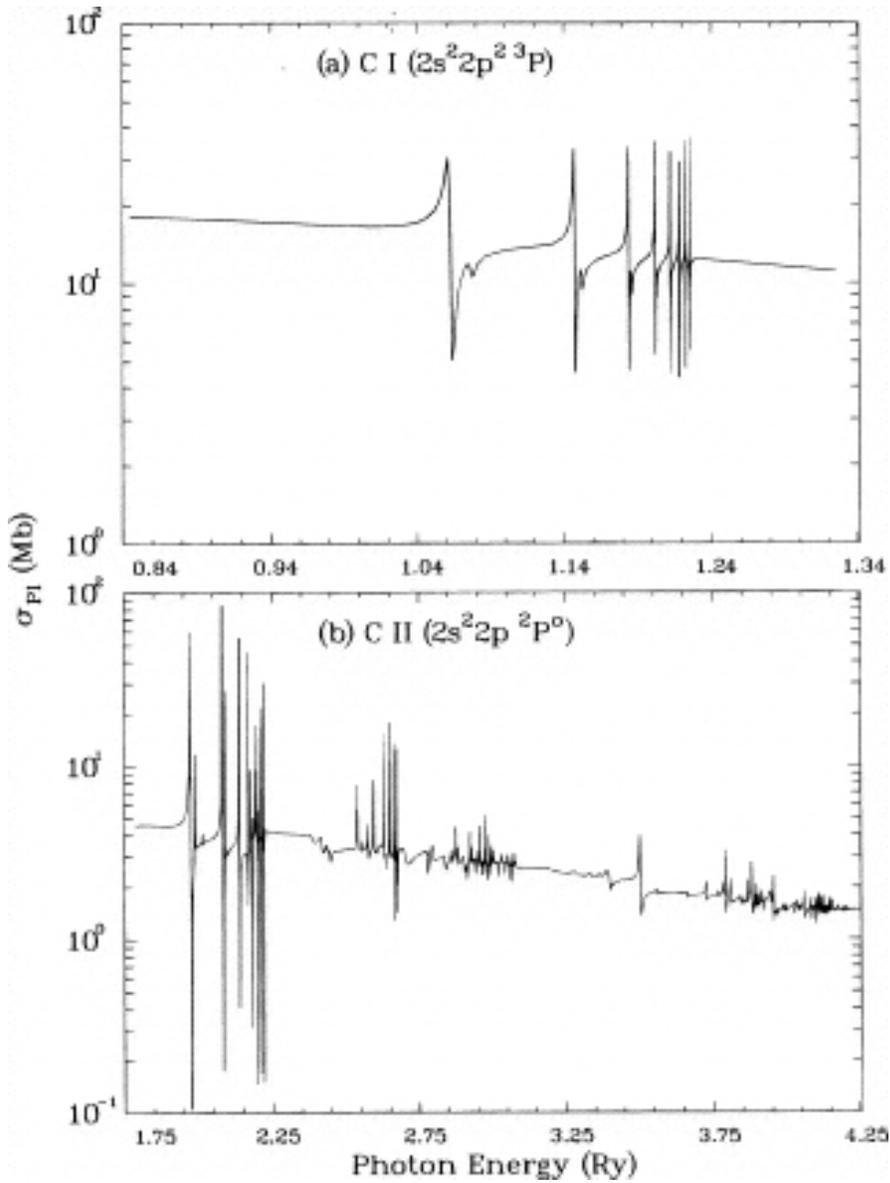
Fig. 5.4. The computed C I and C II photoionisation cross–section in mega–Barns $(10^{-18}\,\mathrm{cm}^2)$ after Nahar and Pradhan (1997). Note the complex resonance structure,

$$a_1 \qquad = 11.132 : a_1 = 5.612 : a_1 = 6.9668 : a_4 = 11.132$$
$$a_5 \qquad = 0.35327 : b_1 = 1.6798 \times 10^{-2} : b_2 = 3.6703 \times 10^{-4} :$$
$$b_3 \qquad = 9.7991 \times 10^{-4} : b_4 = 8.0332 \times 10^{-3}$$

while the threshold energies are given in the range $1 \leq Z \leq 26; 1 \leq N \leq Z$ by:

$$E_{th} = I_H Z^2 N^\gamma : \gamma = 0.20 - 0.39/\log[Z] \qquad (5.41)$$

5.4 Charge–Exchange

During the collision of two ionic species, the charge clouds surrounding each interact, and it is possible that an electron is exchanged between them. Since, in virtually all diffuse astrophysical plasmas, hydrogen and helium are over-whelmingly the most abundant species, the charge–exchange reactions which are significant to the ionisation balance of the plasma are:

$$A^{(i+1)+} + H^0 \;\rightleftharpoons\; A^{i+} + H^+ + \Delta E \text{ and}$$
$$A^{(i+1)+} + He^0 \rightleftharpoons A^{i+} + He^+ + \Delta E$$
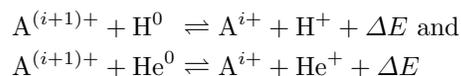
The reactions are exothermic, because of the lower ionisation potential of the A^i ion, and therefore there is an energy barrier against the reverse reaction. This is the energy needed to push the two charged species to within a distance at which charge–exchange can occur in the presence of the Coulomb barrier. Thus, the reverse reaction channel only opens up when $kT_e \gtrsim \Delta E$. In many cases we only have to consider the forward reaction.

Charge–exchange may also occur in collisions of molecules with atoms, *i.e.*

$$CO^+ + H^0 \rightarrow CO + H^+ \qquad (5.42)$$

this process is also closely related to *ion–molecule reactions* such as:

$$CH^+ + O^0 \rightarrow CO + H^+ \qquad (5.43)$$

and to *associative detachment reactions* such as

$$O^- + O \rightarrow O_2 + e^- \qquad (5.44)$$

These and other molecular reactions are dicussed in Chapter 13.

Advanced Section:

During the collision, between an ion $A^{(i+1)}$ and a hydrogen or helium atom, we can regard a charged molecule of the form $HA^{(i+1)}$ or $HeA^{(i+1)}$ to have been effectively formed, albeit for a rather short time. In the $A^{(i+1)} + H^0$ channel, the interaction potential as a function of inter–nuclear distance, r, results from the polarisability of the H^0 or He^0 atom, α_{pol}:
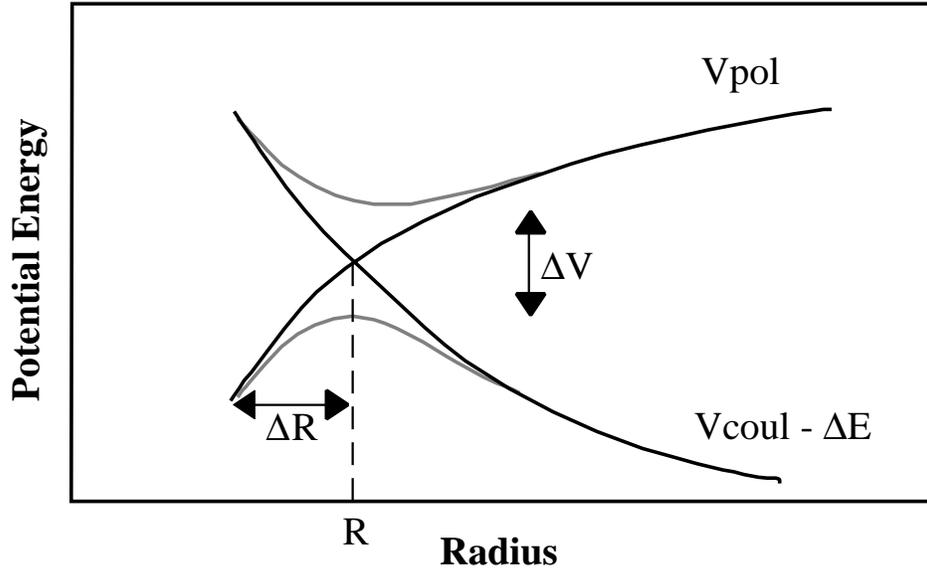
Fig. 5.5. Potential states formed in a charge–exchange reaction.

$$V_{pol}(r) = -\frac{\alpha_{pol}}{2}\left(\frac{e}{r}\right)^4 \tag{5.45}$$

where $\alpha_{pol}(\mathrm{H}) = 6.7 \times 10^{-25}\,\mathrm{cm}^2$ and $\alpha_{pol}(\mathrm{He}) = 2.07 \times 10^{-25}\,\mathrm{cm}^2$.

In the $\mathrm{A}^{i+} + \mathrm{H}^+$ channel the Coulomb interaction determines the long–range interaction potential:

$$V_{coul}(r) = i\left(\frac{e}{r}\right)^2 \tag{5.46}$$

These two potential curves cross at an inter–nuclear distance, R, determined by $V_{coul}(R) - \Delta E = V_{pol}(R)$. When $V_{pol}(R) \ll \Delta E$, which is usually the case, then $R \sim ie^2/\Delta E$.

Two classes of charg–exchange interaction are possible:

- Those in which the collision timescale, t_{coll}, is long enough that the electrons have a chance to continually adjust to the changing inter–nuclear distance. In this case, the interaction potential curves do not cross, but instead follow a form sketched by the dotted lines in Figure (5.5). Such an interaction is *adiabatic*.
- Those (high impact energy) interactions in which the collision timescale is too short for electronic orbital adjustment to occur. These are called *diabatic* interactions.

The collision timescale is determined by the period in which the atoms are within their interaction radius; $t_{coll} \sim \Delta R/v(R)$, while the timescale

for adjustment of the electrons is determined by the Heisenberg Uncertainty Principle; $t_{elec} \sim h/\Delta V$.

In the case of *adiabatic* collisions, $t_{coll}/t_{elec} \gg 1$, the ions are effectively in orbit about one another, and the critical impact parameter (projected internuclear distance in the initial direction of impact), for any impact energy E is determined from energetic considerations:

$$r_{12} = \left(\frac{2e^2 \alpha_{pol}}{E} \right)^{1/4} \tag{5.47}$$

During the collision, the electron will lose all memory of which channel it approached in, and so it is equally likely to emerge in either channel. The probability of charge–exchange is therefore $1/2$, and the effective cross section for charge–exchange is therefore $\sigma_{ce} = \pi r_{12}^2/2$. The rate constant is the product of the cross–section and the velocity $v = (2E/\mu)^{1/2}$: where $\mu = m_1 m_2/(m_1 + m_2)$ is the reduced mass of the reactants;

$$\alpha_{ce} = \langle \sigma_{ce} v \rangle = \pi e \left(\frac{\alpha_{pol}}{\mu} \right)^{1/2} \tag{5.48}$$

$$\sim 1.5 \times 10^{-9} \, \mathrm{cm^3 \, s^{-1}} \text{ for c.e. with H}$$

$$\sim 8 \times 10^{-10} \, \mathrm{cm^3 \, s^{-1}} \text{ for c.e. with He.}$$

In the *Landau–Zenner* case of *diabatic* charge–exchange the rate is summed over the contributions arising from the different orbital angular momentum quantum numbers. Since more channels open up as energy increases, the charge–exchange rate initially increases rapidly with temperature, but with increasing impact energy, the probability of charge exchange decreases, and the charge–exchange rate falls. This behaviour is illustrated in Figure (5.6).

Normally, the charge–exchange reaction will leave the ion in the lowest available energy state. However, provided that this charge–exchange reaction is sufficiently exothermic, then charge–exchange which leaves the ion in various excited states becomes energetically possible, increasing the total charge–exchange reaction rate. This is illustrated in the case of Si^{3+} in its charge–exchange reactions with H:

$$Si^{3+}(3s \; {}^2S) + H \rightarrow \; Si^{2+}(3s2p \; {}^1P^0) + H^+ + 9.62eV$$
$$Si^{3+}(3s \; {}^2S) + H \rightarrow \; Si^{2+}(3p^2 \; {}^1D) + H^+ + 4.74eV$$
$$Si^{3+}(3s \; {}^2S) + H \rightarrow \; Si^{2+}(3p^2 \; {}^3P^0) + H^+ + 3.80eV$$
$$Si^{3+}(3s \; {}^2S) + H \rightarrow \; Si^{2+}(3s3d \; {}^3P^0) + H^+ + 2.17eV$$

and in its charge–exchange reactions with He:

$$Si^{3+}(3s \; {}^2S) + He \rightarrow \; Si^{2+}(3s^2 \; {}^1S) + He^+ + 8.88eV$$
$$Si^{3+}(3s \; {}^2S) + He \rightarrow \; Si^{2+}(3s3p \; {}^3P) + He^+ + 2.30eV$$
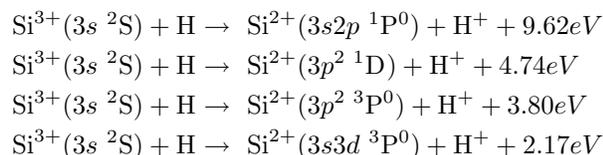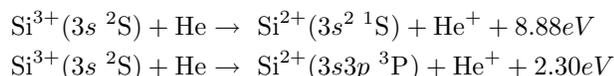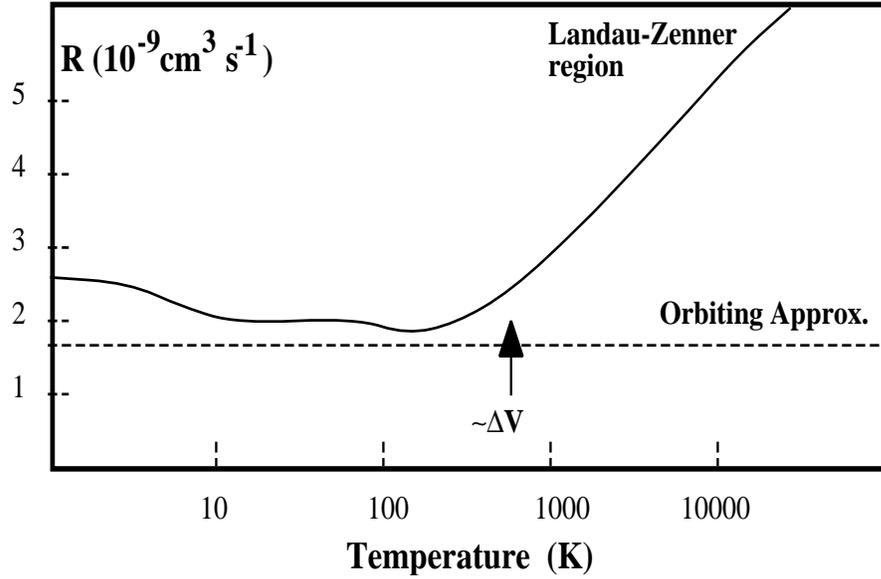
Fig. 5.6. Rate of the $Si^{2+}+H^0 \rightarrow Si^++H^+$ charge exchange reaction as a function of temperature., compared with the expectations of the orbiting approximation (after McCarroll & Valiron, 1976).

5.5 Coronal Equilibrium

Collisional ionisation balance involves, in general, three ionisation stages of a given element, $A^{(i-1)+}$, A^{i+}, and $A^{(i+1)+}$, coupled through collisional ionisation, the various means of recombination and through charge–exchange reactions. If the number density of these species are $n_{A,i-1}$, $n_{A,i}$, and $n_{A,i+1}$, respectively , then the collisional ionisation equilibrium condition is:

$$n_e n_{A,i} \alpha_{coll}^{A,i} + n_{A,i} \sum_{x=H^0,He^0} n_x \alpha_{ce}^{A,i}$$

$$= n_e n_{A,i+1} \alpha_{rec}^{A,i+1} + n_{A,i+1} \sum_{x=H^0,He^0} n_x \alpha_{ce}^{A,i+1} \qquad (5.49)$$

Here, for all except the neutral species, Coulomb interactions ensure that the dominant path for charge–exchange reactions is towards a decrease in the ionisation state of the atom (*i.e.* reactions of the type $H^0 + A^{i+} \rightarrow H^+ + A^{(i-1)+}$, or the corresponding reaction with helium). In equation (5.49), therefore we have therefore ignored charge–exchange reactions which work in the opposite sense, and so have eliminated the coupling to the $(i-1)$ stage of ionisation.

Implicit in this equation (*the coronal approximation*) is the assumption that the ionization rates on the left–hand side refer to the ground state of the

ion. This will only be true provided that radiative and collisional processes between levels in an ion proceed on much faster timescales than photoionisation and recombination. This certainly applies in the limit of low densities, but breaks down at higher densities when some excited states may become significantly populated in their approach towards LTE conditions.

If charge–exchange reactions can be ignored entirely (which would be true for hydrogen or helium in a plasma with solar abundances) then equation (5.49) simplifies marvellously, to give the ionisation balance directly:

$$\frac{n_{A,i+1}}{n_{A,i}} = \frac{\alpha_{coll}^{A,i}}{\alpha_{rec}^{A,i+1}} \tag{5.50}$$

An elegant and simple result indeed!

5.5.1 The Case of a Pure Hydrogen Plasma

The case of a pure hydrogen plasma provides insight into the behaviour of collisional ionised plasmas in general. Here, of course, the electrons are provided only by the ionisations, so that the coupling between the electrons and ions is closer than in other plasmas in which the ion is simply a trace element. In this case the hydrogen ionisation balance is:

$$\frac{n_{H^+}}{n_{H^0}} = \frac{\alpha_{coll}^{H^0}}{\alpha_{rec}^{H^+}} = F(T) \tag{5.51}$$

and the fractional ionisation of hydrogen, χ_H, can therefore be expressed in terms of the function of temperature $F(T_e)$:

$$\chi_H = \frac{n_{H^+}}{n_H} = \frac{n_{H^+}}{(n_{H^0} + n_{H^+})} = \frac{F(T)}{[1 + F(T)]} \tag{5.52}$$

Now, an analytic approximation to $R_{coll}^{H^0}$ has already been given above, in equation (5.10), and a good approximation to $\alpha_{rec}^{H^+}$ was given following equation (5.18); $\alpha_{rec}^{H^+} = 4.18 \times 10^{-13} \left[T/10^4 K\right]^{-0.72}$, from which:

$$F(T_e) = 0.788T_e^{1.22} \left(1 + T_e/78945\right) \exp\left[-157890/T_e\right] \tag{5.53}$$

The fractional ionisation of hydrogen is given in Table 5.1. In an optically–thin plasma hydrogen is 50% ionised at $T_e \sim 14000K$ and is 98% ionised by 20000K. At lower temperatures, the ionisation fraction falls away in a precipitate manner as a result of the exponential factor in eqn (5.53). Note that the ionisation temperature, defined as the temperature where the ion reaches 50% ionisation is very much lower than the ionisation potential energy expressed as a temperature (157890 K). This is the case for heavier elements as well.

Table 5.1. Fractional ionisation of H as a function of the electron temperature

$\log[T_e(K)]$	3.7	3.8	3.9	4.0	4.1	4.2	4.3	4.4
$\log[\chi_H]$	-9.300	-6.372	-4.024	-2.143	-0.716	-0.098	-0.011	-0.002
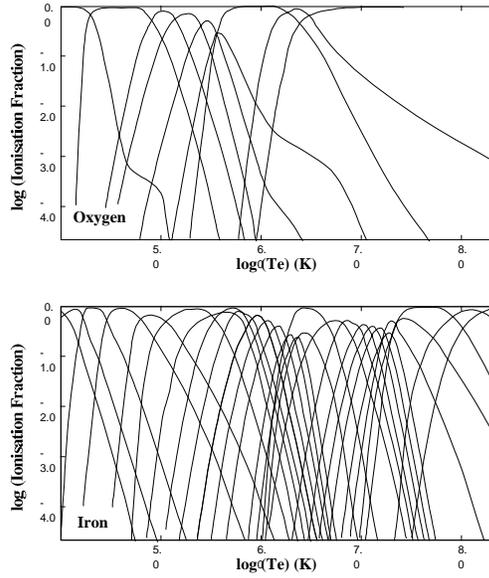
Fig. 5.7. The collisional ionisation equilibrium balance of oxygen and iron as a function of temperature. Note the persistence of hydrogen–like and helium–like species to very high temperatures, and the way in which the ionisation balance echoes the shell structure of the atoms.

5.5.2 Ionisation Equilibrium of Heavy Elements

The ionisation balance of heavy elements can be computed with reasonable accuracy with equation (5.50). However, in practice we solve the full collisional ionisation balance equations (5.49). Figure (5.7) shows the full ionisation balance for the ions of oxygen and of iron taken from the detailed calculations of Sutherland & Dopita (1993).

Note the persistence of the hydrogen–like and helium–like ions to very high temperature. This is because of the difficulty of removing the inner K–shell electrons due to their very high ionisation potentials (739 and 871 eV, respectively in the case of oxygen, compared with 138 eV for the ionisation potential of the Lithium–like ion). In the case of iron, similar behaviour can also be seen for the L–shell electrons at the ion Fe XVIII where the ionisation potential jumps from 489 eV to 1.4 keV.

Recalling the difference in the shape of the collisional ionisation and collisional excitation cross–sections near the threshold energy (*c.f.* Figures **??** and **??**), it is clear that the H– and He– like Lyα lines will have relatively large collisional excitation cross–sections compared with the collisional ionisation cross–sections near the threshold energy for ionisation. As a consequence, each ion will be collisionally excited several times before it is collisionally ionised. This fact makes the Lyα lines of H– and He–like species very important as X–ray cooling lines.

Charge–exchange plays a particularly important role in determining the O I/ O II ionisation balance. This is because the ionisation potential of the O I is 13.618 eV, compared with 13.598 eV for hydrogen. As a consequence the charge–exchange reaction:

$$O^+ + H^0 \rightleftharpoons O^0 + H^+ + 0.02eV \qquad (5.54)$$

is almost resonant, and the reverse reaction is opened up for temperatures above about 1000 K. However, hydrogen becomes ionised at about 14000 K, as we demonstrated above. In effect, the oxygen ionisation balance is locked to that of hydrogen in the temperature range where are both partially ionised.

Notes on Chapter 5

- The best source for analytic fits to the collisional ionisation and excitation autoionisation cross–sections is that of Arnaud, M. & Rothenflug, 1985, A&AS, 60, 425.
- Auger ionisation data is given by Kaastra, J.S. & Mewe, R. 1993, A&AS, 97, 443.

Exercise 5.5.1. The ionisation balance of nitrogen and oxygen (at least as far as the un-ionized and singly-ionized states are concerned) is locked to that of hydrogen by the (nearly resonant) charge exchange reactions:

$$O^+ + H^0 \rightleftharpoons O^0 + H^+ + 0.019eV$$
$$N^+ + H^0 \rightleftharpoons N^0 + H^+ + 0.935eV$$
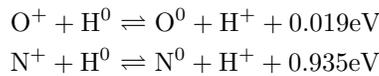
Because these reactions are exothermic by a small amount of energy ΔE, the rate of the forward reaction is higher than the rate of the reverse reaction by the ratio $\frac{g1}{g2} \exp[\Delta E/kT]$ where the reatio of the statistical weights, $\frac{g1}{g2} = \frac{9}{8}$ for the oxygen charge-exchange reaction, and $\frac{g1}{g2} = \frac{2}{9}$ for the nitrogen charge exchange reaction. Assuming that hydrogen is in collisional ionization equilibrium as given in Table (5.1), calculate the equilibrium ionization fractions of oxygen and nitrogen at the same temperatures as in the Table.

Exercise 5.5.2. Using the formulae given in section (5.3.4), calculate the K-shell photoionization threshold energies and the photoionization cross sections just above these energies for O I through O VIII. For the case of O VIII, how well do equations (5.37) and (5.39) agree?

Exercise 5.5.3. (for advanced students) Using any technique (Fortran, Pascal or C program, Maple, Mathematica or just pen, pencil and brain) answer the following questions. **Hint:** read section (5.5):

a. If a cloud of plasma in a supernova remnant is at in collisional ionization equilibrium at a temperature of 250,000K, then what are the fractions of each of the carbon ions present?

b. In collisional ionization equilibrium, at what temperature would Carbon IV be the most abundant ion? (solve graphically in the temperature range $3 \times 10^4 - 3 \times 10^4$).

Data for the problem:

Use the following simplified formulae for the required rates (from Landini & Monsignori Fosse, 1990, A&AS, 82, 229):

Collisional rates:

$$\alpha_{\rm c} = A_{\rm col} \frac{T^{1/2}}{(1 + 0.1 T/T_{\rm col})} \exp\left[-\frac{T_{\rm col}}{T}\right]$$

Radiative Recombination Rates:

$$\alpha_{\rm r} = A_{\rm rad} \left[\frac{T}{10^4 {\rm K}}\right]^{-X_{\rm rad}}$$

Di-electronic Recombination Rates:

$$\alpha_{\rm d} = A_{\rm di} T^{-3/2} \exp\left[\frac{T_0}{T}\right] \left(1 + B_{\rm di} \exp\left[-\frac{T_{\rm i}}{T}\right]\right)$$

Atomic Data for Carbon:

Ion	$A_{\rm col}$	$T_{\rm col}$	$A_{\rm rad}$	$X_{\rm rad}$	$A_{\rm di}$	$B_{\rm di}$	T_0	$T_{\rm i}$
I	1.44E-10	1.31E5	0	0	0	0	0	0
II	4.20E-11	2.83E5	4.7E-13	0.624	2.54E-3	4.42E-2	1.57E5	3.74E5
III	1.92E-11	5.56E5	2.3E-12	0.645	6.15E-3	5.88E-2	1.41E5	1.41E5
IV	5.32E-11	7.48E5	3.2E-12	0.770	1.62E-3	3.43E-1	8.19E4	1.59E5
V	2.87E-13	4.55E6	4.9E-12	0.803	4.78E-2	3.62E-1	3.44E6	5.87E5
VI	9.16E-14	5.68E6	9.2E-12	0.791	3.22E-2	3.15E-1	4.06E6	8.31E5
VII	0	0	3.2e-11	0.718	0	0	0	0

6. Continuum & Recombination Line Processes

"I ask you to look both ways.
For the road to a knowledge of the stars leads through the atom;
and important knowledge of the atom has been reached through the
stars"
— *Sir Arthur Eddington (Stars & Atoms, 1928)*

Transitions of electrons between discrete levels of an atomic or molecular system can produce photons of specific energies. These nearly monochromatic photons form the by now familiar line emission from and absorption by astrophysical plasmas. The discrete levels as also known as bound levels and they arise in the quantum mechanical description as quantised solutions to electrons bound to a nucleus in some form of potential. Outside the potential electrons are unbound or free. The free electrons can take on any energy and can occupy any of a continuum of states. Nevertheless, electrons can make transitions between these free states, known as free–free transitions, and between bound atomic levels and free states, known as bound–free or free–bound transitions. Using this nomenclature then the line transitions discussed so far are bound–bound transitions.

Free–free and free–bound transitions can emit photons of a whole range of energies, and tend to form a 'continuum' or continuous emission, providing a background on which the lines appear in a spectrum.

6.1 Free–Free Continuum Emission

In a classical result first established by Larmor in 1897, any charge which is accelerated in space will emit electromagnetic radiation. For an electron, the total power radiated, P, is:

$$P = \frac{2e^2}{3c^3}\left(\frac{dv}{dt}\right)^2 \tag{6.1}$$

Such accelerations occur in ionised plasmas as a result of the collision between an electron and a proton or other ion in the plasma, and this gives a mechanical description of the basis of free–free emission. Taking an ensemble of particles with thermal velocities and cross–sections a model which
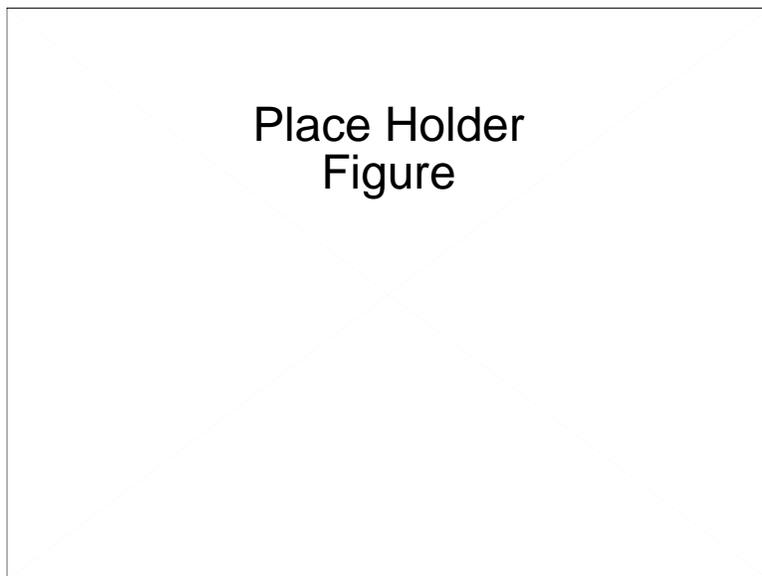
Fig. 6.1. Spectrum Showing Lines and Continuum

works out the distribution of deflections experienced and corresponding ac-celerations could be used to predict a spectrum. For example, the classical cross–section for free–free absorption is given by Kramers (1923),

$$d\sigma_{class} = a_0 Z^2 \frac{32\pi}{3\sqrt{3}} \left(\frac{e^2}{\hbar c}\right) \left[\left(\frac{Ry}{h\nu}\right)^3 \frac{d(h\nu)}{Ry}\right]. \qquad (6.2)$$

However,a quantum wave mechanical description can also provide a spectrum, with the added advantage that the wave description correctly treats the inter-actions on the atomic scale where simple mechanics fail. Atoms and electrons, protons and other ions are not well treated as particles on the smallest scales and this changes the 'collisions' and simple application of equation [] will not suffice. Finally, we can use techniques for free–free and free–bound transi-tions that have been used for bound–bound transitions, largely by changing the functional form of the wavefunctions. Instead of a real bound spherical harmonic wavefunction, the free states can be considered as infinitely oscil-lating sinusoidal plane wave. Infinitely oscillating waves are often represented in complex mathematical terms with imaginary terms.

   The approach generally used in all kinds of continuum emission calcu-lations is to use the classical result computed on the basis of straight me-chanics, charge separations and accelerations, and then to work out the full wave–mechanical solution and express the difference as a correction factor to the classical result. This approach was first make popular for continuum

by Gaunt (1930) and these 'correction' factors are known as Kramers–Gaunt factors or more simply as Gaunt factors. When the classical result is accurate these factors are approximately 1.0 and can vary above and below this as quantum mechanics becomes important. Elsewhere in plasma and atomic physics other correction (fudge!) factors used to modify results, such as collision rates, are also called gaunt factors (but usually with a small g). These gaunt factors are generally used to encompass more difficult physics which may be difficult or impossible to compute exactly and allow that to be at least allowed for without unduly complicating matters.

Finally, Gaunt factors for given transitions can be integrated and averaged in various ways for example the free–free Gaunt factor refers to state and electron energies only. To get a Gaunt factor that can be used on a plasma with a Thermal Maxwellian velocity distribution, the Gaunt factor can be averaged over the energy distribution to get a thermal average gaunt factor. This can in turn be integrated over all frequencies to get a gaunt factor useful for correcting the classical expression for the total power emitted by a thermal plasma in free–free emission for example.

Note: this section is pretty symbol heavy. To streamline it and reduce the 'where x is ...' count, a glossary of symbols are collected at the end of the chapter.

6.1.1 free–free Gaunt factors

An electron of initial energy $E_i = \hbar^2 k_i^2/2m_e$ has a cross–section for absorption of a photon of energy $h\nu$ to a final energy $E_f = \hbar^2 k_f^2/2m_e$ which is related to the equivalent cross–section for emission of a photon $h\nu$ from a electron of energy E_f. Hence,

$$\sigma_{i,\nu-f} = \frac{v_f}{c} \frac{\rho_f}{\rho_{i,\nu}} \sigma_{f-i,\nu}, \tag{6.3}$$

where v_f is the electron velocity and ρ_i and ρ_f are the state or continuum level densities:

$$\rho_f = \frac{m_e k_f}{\pi^2 \hbar^2}, \tag{6.4}$$

$$\rho_i = \frac{4\nu^2}{\hbar c^2} \frac{m_e k_i}{\pi^2 \hbar^2}, \tag{6.5}$$

with the free electron wave functions normalised to plane waves of unit density at infinity. This means that emission and absorption are related processes and that if we work out one process we can infer the other. We can begin by looking at absorption and know that the result can be equally applied to emission.

Using overlap integrals $\tau_{l',l}$ between the two electron wavefunctions, the differential cross–section of absorption according to wave mechanics is (Karsas & Latter 1961):

$$d\sigma^{ab}_{i,\nu-f} = a_0^2 Z^2 \frac{64}{3} \left(\frac{\mathbf{e^2}}{\hbar c}\right) \left[\left(\frac{Ry}{h\nu}\right)^3 \frac{d(h\nu)}{Ry}\right] \left[\frac{2\pi^2\hbar^2}{m_e k_i}\right]$$

$$\times k_i k_f \sum_{l=0}^{\infty} [(l+1)\tau^2_{l+1,l} + l\tau^2_{l-1,l}], \qquad (6.6)$$

where $\tau_{l',l}$ is the radial integral of the wave-functions $\Psi_{l'}$ and Ψ_l:

$$\tau_{l',l} = \int_0^{\infty} r^2 \Psi_{l'}(r,E_i) \frac{1}{r^2} \Psi_l(r,E_f) dr, \qquad (6.7)$$

and $\Psi_l \sim \sin(kr + \delta_l)/kr$ asymptotically.

To compute the effective transition probabilities between all the continuum of free states we can again use hypergeometric functions as in Chapter 2 §?? to perform the overlap integrals. Now, because the wave–functions are unbound sinusoidal waves, the hypergeometric functions here (I_l, G_l and $_2F_1$) have *complex* arguments rather than the real arguments used in Chapter 2 when computing single transition probabilities in bound–bound transitions of hydrogen. The sums here are also more complicated because we are summing over infinite waves rather than finite bound ones. The infinite sum Σ in equation (6.6) may be expressed in terms of variables $\eta_i = Z/(a_0 k_i)$ and $\eta_f = Z/(a_0 k_f)$ and hypergeometric functions,

$$\Sigma = \frac{I_0}{k_i k_f} [(k_i^2 + k_f^2 + 2k_f^2\eta_f^2) \frac{I_0}{k_i k_f} - 2k_i k_f (1+\eta_i^2)^{1/2} (1+\eta_f^2)^{1/2} \frac{I_1}{k_i k_f}], \quad (6.8)$$

with

$$I_l = \frac{1}{4} \left[\frac{4k_i k_f}{(k_i - k_f)^2}\right]^{l+1} \exp\left[\frac{\pi}{2}|\eta_i - \eta_f|\right] \frac{|\Gamma(l+1+i\eta_i)\Gamma(l+1+i\eta_f)|}{\Gamma(2l+1)} G_l,$$
$$(6.9)$$

and

$$G_l = \left|\frac{k_f - k_i}{k_f + k_i}\right|^{i\eta_i + i\eta_f} {}_2F_1\left[l+1-i\eta_f, l+1-i\eta_i, 2l+2; -\frac{4k_i k_f}{(k_i - k_f)^2}\right],$$
$$(6.10)$$

with a complex exponent in the first term and complex values for the first two arguments of $_2F_1$.

Finally, after the cross–section is converted to the right units to compare with the classical result we get:

$$d\sigma_{e,i,\nu-f} = a_0^2 Z^2 \frac{64}{3} \left(\frac{\mathbf{e^2}}{\hbar c}\right) \left[\left(\frac{Ry}{h\nu}\right)^3 \frac{d(h\nu)}{Ry}\right]$$

$$\times \frac{I_0}{\eta_i \eta_f} [(\eta_i^2 + \eta_f^2 + 2\eta_i^2\eta_f^2) I_0$$

$$- 2\eta_i \eta_f (1+\eta_i^2)^{1/2} (1+\eta_f^2)^{1/2} I_1]. \qquad (6.11)$$

The ratio of this to equation (**??**) is the correction factor, or Gaunt factor (Gaunt 1930) for free–free transitions. In uniform η variables, it is:

$$g_{ff}(\eta_i, \eta_f) = \frac{2\sqrt{3}}{\pi \eta_i \eta_f} I_0[(\eta_i^2 + \eta_f^2 + 2\eta_i^2 \eta_f^2)I_0 - 2\eta_i \eta_f (1 + \eta_i^2)^{1/2}(1 + \eta_f^2)^{1/2} I_1],$$

(6.12)

where the ηs are also related to the electron energies by $\eta_i^2 = Z^2 Ry/E_i$, $\eta_f^2 = Z^2 Ry/E_f$.

Normalising units, working in scaled electron energy, $\varepsilon_i = E_i/Z^2 Ry$, and scaled photon energy, $w = h\nu/Z^2 Ry$, with $\eta_i = 1/\varepsilon_i^{1/2}$ and $\eta_f = 1/(\varepsilon_i+w)^{1/2}$, equation [6.12] was then evaluated, giving $g_{ff}(\varepsilon_i, w)$ over the ranges $10^{-8} < \varepsilon < 10^9$ and $10^{-8} < w < 10^9$ in figures **??** & **??** and in abbreviated form in Table **??**.

Temperature Averaged Gaunt Factors. Assuming a Maxwell–Boltzmann electron distribution, it is a straightforward calculation to integrate $g_{ff}(\varepsilon_i, w)$ over the electron distribution, for a range of photon energies.

Changing variables and integrating we get the temperature averaged free–free gaunt factor

$$< g_{ff}(\gamma^2, u) > = \int_0^\infty e^{-x} g_{ff}\left(\varepsilon_i = \frac{\sqrt{x}}{\gamma}, w = \frac{u}{\gamma^2}\right) dx,$$

(6.13)

where $\gamma^2 = Z^2 Ry/kT$ and $u = h\nu/kT$.

This factor is plotted in figures **??** to **??**.

Free–Free Emissivity. By summing over all ions present in a plasma, the total free–free emission from the plasma can be evaluated at a given frequency and temperature using

$$P_{ff}(\nu, T) = n_e n_H f_k \frac{\exp(-u)}{\sqrt{T}} \sum_{el,Z} A_{el}\chi_{el,Z}Z^2 < g_{ff}(\gamma^2, u) > \text{ erg cm}^{-3}\text{s}^{-1},$$

(6.14)

where f_k is the collection of constants:

$$f_k = \frac{16}{3\sqrt{3}} \left(\frac{\pi}{2km_e^3}\right)^{1/2} \left(\frac{e^2}{c}\right)^3,$$

(6.15)

and has a numerical value of 5.44436×10^{-39} using modern constants in c.g.s. units. The variables n_e and n_H are the number densities of electron and hydrogen atoms respectively, and A_{el} is the abundance of the element el relative to hydrogen by number, and $\chi_{el,Z}$ is the ionisation fraction of the ions of element el in stage Z.

Total Free–Free Emission Gaunt Factors. Integration of the temperature averaged Gaunt factor over the photon spectrum, in u yields the total energy emission Gaunt factor,
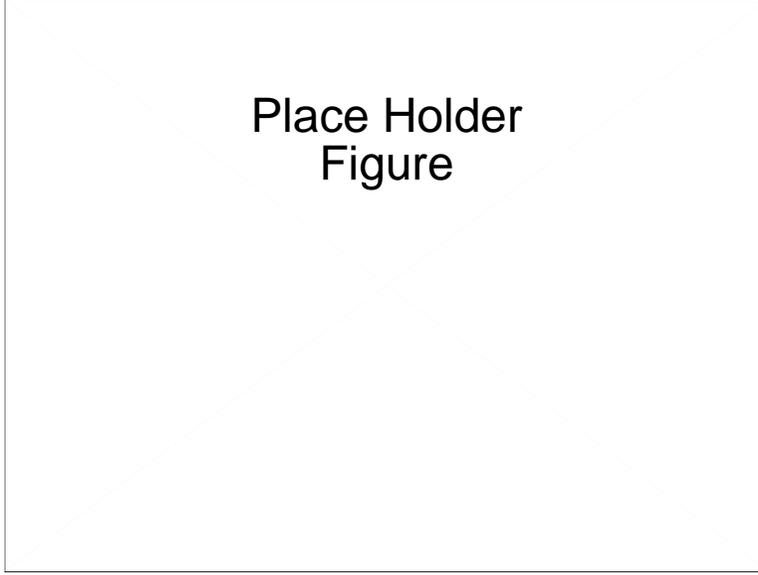
Fig. 6.2. $\log - \log$ plot of the integrated free–free Gaunt factor $< g_f f(\gamma^2) >$ as a function of $\gamma^2 = Z^2 Ry/kT$.

$$< g_{ff}(\gamma^2) >= \int_0^\infty \exp(-u) < g_{ff}(\gamma^2, u) > du. \qquad (6.16)$$

The total free–free emission due to electrons at a given temperature, from a plasma of many ions is,

$$\Lambda_{ff} = n_e n_H F_k \sqrt{T} \sum_{el,Z} A_{el}\chi_{el,Z}Z^2 < g_{ff}(\gamma^2_{eff}) > \text{ergs cm}^{-3} \text{ s}^{-1} \qquad (6.17)$$

where $\gamma^2_{eff} = I_{el,Z}/kT$ is the scaled ionisation potential of element el in ionisation stage Z for each ion (serving the same function as $Z^2 Ry$), and the collection of constants,

$$F_k = \frac{16}{3\sqrt{3}} \left(\frac{2\pi k}{\hbar^2 m_e^3} \right)^{1/2} \left(\frac{\mathbf{e}^2}{c} \right)^3, \qquad (6.18)$$

has the numerical value of 1.42554×10^{-27} in c.g.s units.

6.2 The Free–Bound Continuum

Free–bound transitions is just another way of describing the recombination process. Since the free electron can have a range of energy, usually in a thermal
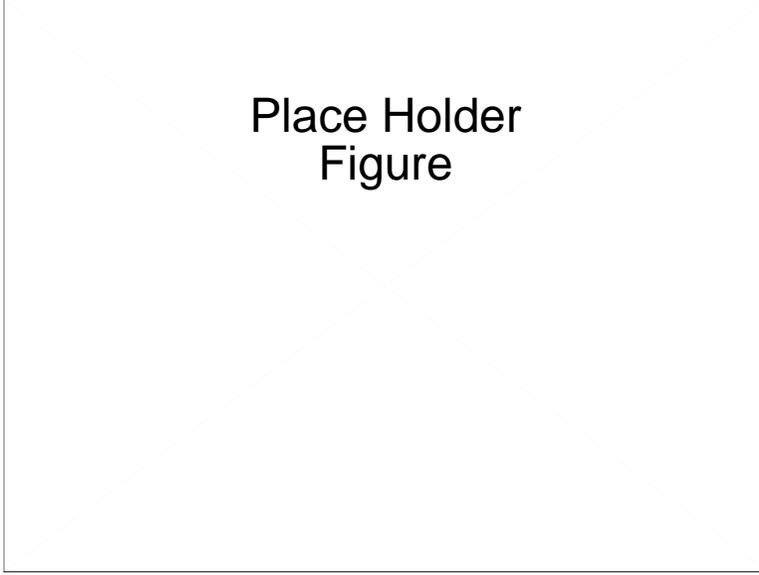
Fig. 6.3. Computed free–free emission efficiencies $\chi_{ff} = \Lambda_{ff}/\Lambda_{\text{tot.}}$, for power–law photoionised plasmas. The solid curve represents a series of solar metallicity models and the dashed curve is for a primordial H and He composition. Free–free emission can account for as much as 10-20% of the total emission in extreme solar metallicity models, and more in metal poor models.

Table 6.1. $< g_{ff}(\gamma^2, u) >$

	$\log(u = h\nu/kT)$				
$\log(\gamma^2)$	-4.0	-2.0	0.0	2.0	4.0
-4.0	5.5243(0)	5.4983(0)	5.0090(0)	3.8317(0)	2.7008(0)
-3.0	4.2581(0)	4.2403(0)	3.7816(0)	2.7008(0)	1.8041(0)
-2.0	3.0048(0)	3.0152(0)	2.6560(0)	1.8071(0)	1.2771(0)
-1.0	1.8153(0)	1.8880(0)	1.7825(0)	1.2886(0)	1.0747(0)
0.0	8.5314(-1)	9.6975(-1)	1.2939(0)	1.1033(0)	1.0237(0)
1.0	3.1012(-1)	3.9000(-1)	9.7254(-1)	1.0825(0)	1.0202(0)
2.0	1.0069(-1)	1.3352(-1)	5.1714(-1)	1.1065(0)	1.0355(0)
3.0	3.1977(-2)	4.3211(-2)	1.9971(-1)	9.5479(-1)	1.0680(0)
4.0	1.0121(-2)	1.3760(-2)	6.7498(-2)	5.1462(-1)	1.1040(0)

Table 6.2. $< g_{ff}(\gamma^2) >$

$\log(\gamma^2)$	$< g_{ff}(\gamma^2) >$	$\log(\gamma^2)$	$< g_{ff}(\gamma^2) >$
-4.0	1.11388	0.0	1.43220
-3.8	1.11698	0.2	1.41391
-3.6	1.12089	0.4	1.38830
-3.4	1.12581	0.6	1.35832
-3.2	1.13200	0.8	1.32658
-3.0	1.13975	1.0	1.29496
-2.8	1.14945	1.2	1.26462
-2.6	1.16149	1.4	1.23618
-2.4	1.17635	1.6	1.20993
-2.2	1.19447	1.8	1.18594
-2.0	1.21622	2.0	1.16421
-1.8	1.24182	2.2	1.14464
-1.6	1.27104	2.4	1.12711
-1.4	1.30328	2.6	1.11147
-1.2	1.33711	2.8	1.09757
-1.0	1.37040	3.0	1.08526
-0.8	1.40029	3.2	1.07438
-0.6	1.42365	3.4	1.06481
-0.4	1.43768	3.6	1.05640
-0.2	1.44060	3.8	1.04904
0.0	1.43220	4.0	1.04264

distribution, the possible transition energies range from the energy difference between the level energy and the ionisation potential of the ion, where the contunuum begins, up to the highest energy available in the free electrons. In a spectrum this appears as an 'edge' followed by a continuum of emission to higer and higher energies. The shape, or energy dependence, of the continuum is a function of the distribution of electron energies in the continuum and the cross–section for capture as a function of the electron energy. At sufficently high energies both the decreasing availabilty of electrons and the low overlap integral between the free electron wave function and the bound level wave function both serve to cut off the high energy continuum from a given edge.

In a manner analagous to computing the free–free transition gaunt factors, a free–bound gaunt factor can be worked out. In this case the free wavefunction is an imaginary oscillating wave and the bound function is a real function. This is also then similar to the bound–bound transition probability calcualtions of chapter 2. As might be expected the transition gaunt factors can be evalutated using hypergeometric functions when the bound level is a hydrogen like state.

[eqn b-f gaunt factor]

In nebula calculations the continuum is often evaluted at a given frequency (energy) in terms of both the sum of the free–free gaunt factors for each species (each Z) present and the and the sum of all bound–free gaunt factors
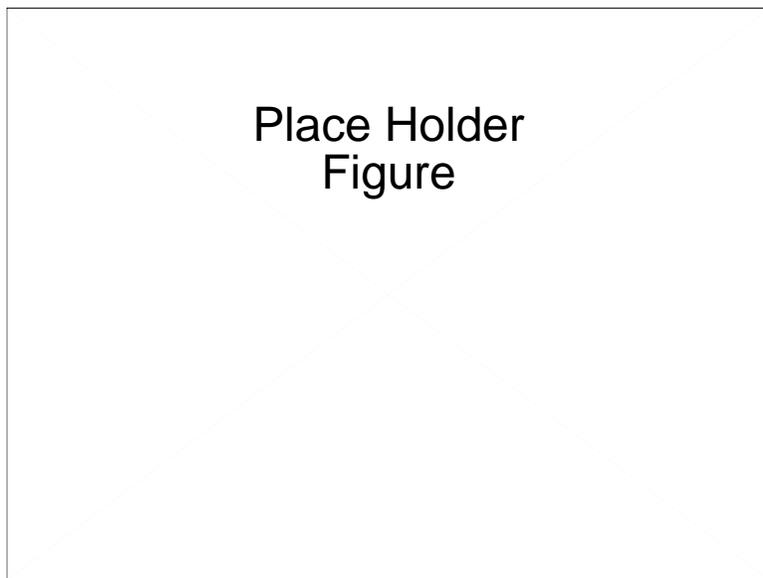
Fig. 6.4. free-bound energy diagram

(one for each edge below the energy in question, for each species present) all of which are added and the net continuum emission for both processes is calculated by:

[eqn free–free + free–bound continuum]

Free–bound gaunt factors for hydrogenic species have been computed accurately (Karsas & Latter 1961; Ferland 1980) and species more complex than hydrogen are sometimes evalutated numerically, using hydrogenic approximations or using experimental data.

6.3 The Two–Photon Process

In the recombination of hydrogen, an appreciable fraction of captured electrons will end up in the $2s$ level. In addition, the absorption of Lyα photons in an ionised nebula will overpopulate the $2s$ level, encouraging the collisional transition to the $2p$ state, rather than the return to the $1s$ state with the re–emission of another Lyα photon. (See Figure **??**) The transition $2p \rightarrow 1s$ is forbidden for dipole radiation according to the selection rules, and although a magnetic dipole transition to the ground state is possible, it turns out that the so–called two–photon process is more probable. The theory of this process was worked out by Breit and Teller (1940) and first applied to ionised nebulae by Spitzer & Greenstein (1951).

The two–photon process can be thought of as a spontaneous dipole transition into a virtual p level between the excited $2s$ state and the ground level, accompanied by a simultaneous transition from this excited state back down to the ground level. Two photons are therefore emitted, each carrying away one quantum unit of angular momentum. The conservation of angular momentum demands that the emission probability is proportional to the square of the cosine of the angle between the polarization vectors.

Consider an intermediate p state lying $\Delta E = h\nu_1$ below the $2s$ level. In this case two photons will be emitted such that $v_1 + \nu_2 = \nu_\alpha$, where $\nu_\alpha = 3cR_H/4h$ is the frequency of the Lyα transition and R_H is the Rydberg constant for hydrogen. From equations (**??**) and (**??**) it is evident that the joint transition probability is:

$$A = A_a A_b \propto \left(\frac{64\pi^4 \mathbf{e}^2}{3hc^3} \right)^2 \nu_1^3 \nu_2^3 \left| F(r) \right|^2 \tag{6.19}$$

where $F(r)$ is a function of the quantum mechanical radial integrals. Note that the probability of emission of a photon of frequency ν will be symmetric about $\nu_\alpha/2$. However, the emissivity is weighted by the energy of the photon, with the result that the emissivity peaks at higher frequency than this, at a wavelength of about 1550Å rather than at 2403Å.

The terms with frequency dependence can be collected into a function $\psi(\nu/\nu_\alpha)$ giving the probability of emission of a photon of frequency ν and a second of frequency $(1 - \nu/\nu_\alpha)\,\nu_\alpha$:

$$dA_{2s1s}(v) = \frac{3}{4} \left(\frac{\pi \mathbf{e}^2}{hc} \right)^6 \psi(\nu/\nu_\alpha)d\nu = 1.7696 \times 10^{-15} \psi(\nu/\nu_\alpha)d\nu \tag{6.20}$$

The total two–photon transition probability is half of the integral of this function over all frequencies, since there are two photons for each transition to the ground state. The function $\psi(\nu/\nu_\alpha)$ is tabulated by Spitzer & Greenstein (1951), who also evaluate the integral

$$\int_0^{\nu_\alpha} \psi(\nu/\nu_\alpha)d(\nu/\nu_\alpha) = 3.770, \tag{6.21}$$

giving the total two–photon transition probability $A_{2s1s} = 8.227\,\mathrm{s}^{-1}$.

Two–photon continuum emission will arise wherever there is a metastable excited s – state. For example in helium, the transition $1s2s\ {}^2S_0 \to 2s^2\ {}^1S$ is forbidden for dipole or multipole transitions, be they electric or magnetic. Nonetheless, these are connected by the two–photon emission with a probability similar to that of hydrogen.

Under certain circumstances the strength of the two–photon continuum can be enhanced relative to the recombination line or continuum emission. As was shown in the previous chapter, the optical depth in the Lyα line may

become very high. Although the transition probability back to the ground state is very high; $A_{2p1s} = 4.68 \times 10^8\,\mathrm{s}^{-1}$, there is a finite probability (1.76×10^{-8}) that in very optically–thick plasmas, the excited electron will make a collisional $2p \to 2s$ transition followed by a two photon transition back to the ground state. If the optical depth in the Lyα can exceed a few 10^7, then this route may lead to the loss of an appreciable fraction of Lyα photons into the two–photon continuum. Such conditions are rare, but may occur for example, in supernova fireballs or in the broad–line regions of active galactic nuclei.

A more common reason for the enhancement of the two–photon continuum is population of the $2s$ state by direct collisions from the ground state, followed by radiative decay through the two–photon continuum. This will occur when the plasma has been shock–excited to a high (but not too high) temperature so that a number of collisional excitations can occur to the $2s$ and $2s$ levels before the hydrogen atom becomes ionised. These conditions are satisfied for, and the effect is very prominent in the low–excitation Herbig–Haro objects – as was shown by Dopita, Binette & Schwartz (1982).

[Table Two Photon Probabilty function]

6.4 Recombination Line Emission

6.4.1 Recombination Line Spectra

The permitted hydrogen and helium lines observed at UV, optical and IR wavelengths in ionised nebulae arise as a result of radiative recombination; see Section (5.2). As we will see in later sections, the recombination line spectrum is fundamental to an understanding of the physics of H II regions. In particular, the ratio of intensities of the forbidden lines of heavier elements to the recombination lines of hydrogen provides us with the principal means of determining chemical abundances in ionised plasmas. Furthermore, since the ratio of recombination line intensities are well–determined from theory, we can use the observed ratios to determine the *reddening*; the degree to which dust has absorbed the light of the nebula in its wavelength–dependent fashion.

Consider a transition from an excited state of hydrogen or helium with principal quantum number m to a lower one with principal quantum number n. In the limit that collisional excitations are unimportant in populating any of the levels, then the population of level n must be entirely due to the sum of direct recombinations from the continuum, and from recombinations into higher levels, followed by radiative cascade leading to the n-th level. The line emissivity per unit volume in the $m \to n$ transition is therefore:

$$F_{mn} = h\nu_{mn} \left(\alpha_m + \sum_{j>m}^{\infty} \alpha_j C_{jm} \right) P_{mn} n_i n_e = h\nu_{mn} \alpha_{mn}^{eff} n_I n_e \qquad (6.22)$$

where n_i is the number density of the ion which is recombining (H^+, He^+ or He^{++}), n_e is the electron density, α_j is the direct recombination rate to excited state, and C_{jm} is the probability that a recombination to level j is followed by a radiative cascade from the j-th to the m-th state over all possible paths. The term in parentheses is therefore the total recombination coefficient into the exited level. P_{mn} is the *branching probability*; that is, the probability that downward transitions from the n-th level proceed through the $m - n$ transition. As equation (6.22) also shows, we can combine all of the physics of the recombination – cascade into a single number α_{mn}^{eff} for each transition, which is the effective recombination rate for the production of a photon in the $m - n$ transition. The trick is to calculate these numbers for any particular temperature, density, and optical depth in the line.

Although, for H^+ and He^{++}, the recombination rates and all the relevant transition probabilities can be calculated exactly from quantum mechanics, there are a number of complicating factors which make the solution of (6.22) non–trivial. First amongst these is the fact that the ionised (H II) region will frequently be optically thick in the Lyman line series.

A more complex problem arises from the fact that the different l –terms of the level with the same principal quantum number n are not populated amongst themselves according to their LTE ratios. We already explained how this comes about in Section (5.2). Cascades from high l states tend to proceed one step at a time through $n, l \rightarrow (n - 1), (l - 1)$ transitions, since these are the only ones permitted by the selection rules. On the other hand, the nP–levels (or the nS–levels with $n > 2$, which have no place other than the nP–levels in which to cascade) are rather efficiently drained towards the ground state, since transitions such as $n, 1 \rightarrow 1, 0$ or $n, 1 \rightarrow 2, 0$ are not only permitted but have very large transition probabilities.

The full n, l recombination – cascade problem must therefore be solved (*i.e.* Brocklehurst, 1971; Brocklehurst & Seaton, 1972). The branching probability for the $n_1, l_1 \rightarrow n, l$ transition is then given by:

$$P_{n_1,l_1:n,l} = A_{n_1,l_1:n,l} / \left(\sum_{n_2=n_0}^{n_1-1} \sum_{l_2=l_1\pm1} A_{n_1,l_1:n_2,l_2} \right) \qquad (6.23)$$

where $n_0 = 1$ in Case A and $n_0 = 2$ in Case B, and the total recombination coefficient into the excited n_1, l_1 state is:

$$\alpha_{n_1,l_1}^{tot} = \alpha_{n_1,l_1} + \sum_{n_3>n_1}^{\infty} \sum_{l_3=l_1\pm1} P_{n_3,l_3:n_1,l_1} \alpha_{n_3,l_3}^{tot} \qquad (6.24)$$

With these equations, the full recombination line spectrum can be solved by proceeding from higher to lower levels, provided the sum over the n states is pursued to high enough n. In practice, the $b_{n.l}$ occupation factors of the highest n states tend towards unity, as the transition probabilities are small

enough to allow collisions to redistribute electrons between the different l and n states.

Extensive tables of effective recombination coefficients and of relative line intensities helium– and hydrogen– line series are given as a function of temperature and density in Brocklehurst (1971), Seaton (1978) and in the various references given in the notes, and will not be repeated here. Also Hummer 1987, 1996)

[Appendix]

6.4.2 The Radio Recombination Lines

Advanced Section

In section (5.2), we gave a qualitative discussion of the level occupation factors $b_{n,l}$, applying to recombining atoms. The occupation factor is defined as the ratio of the actual level population to that which would be expected if LTE conditions apply. As a result of radiative captures to excited states, and the radiative cascade down to the ground state, in hydrogen this may produce $b_{n,l}$ factors which differ very much from unity, especially for the states of lower principal quantum number n where even the occupation factors between different l sub–levels may differ strongly both from unity and from each other.

However, a small fraction of recombinations, of order one in a million will occur to states with $n > 50$, and these will produce recombination lines at radio frequencies less than about 50GHz. For these recombinations the radiative transition probabilities are very much lower, and because of the very large electronic orbitals, the collisional rate is very much higher. These are the conditions needed to restore LTE conditions, and states with large enough n, the $b_{n,l}$ factors will tend to return to unity. However, deviations from strict LTE in the excited states which produce radio recombination lines are still sufficient to produce interesting physical effects, which were first described properly by Goldberg (1966).

The mean orbital radius for an atom in an $n > 50$ state is very large, and the central charge will appear point–like. For such a configuration, the energy levels are hydrogenic. In this case, the frequency of a transition between an upper state with principal quantum number m and a lower state with principal quantum number n is:

$$\frac{\nu}{\nu_0} = \left(\frac{1}{m^2} - \frac{1}{n^2} \right) \sim \frac{2(m-n)}{n^3} \tag{6.25}$$

where ν_0 is the ionisation potential of the ion of interest expressed as a frequency (for hydrogen, this is the frequency of the Lyman limit). Usually, only the recombination lines of hydrogen and helium are strong enough to be observable at radio frequencies. We generally refer to the states having $m - n = 1$ as α transitions, those with $m - n = 2$ as β transitions by

analogy with the corresponding nomenclature for low-n transitions such as Lyα, Lyβ or Paα, Paβ *etc.* Likewise, since these UV, optical or IR series are united by a common lower state ($n = 1$ for the Lyman series; $n = 3$ for the Paschen series) we identify the particular transition by the atom and the m quantum number. For example H40α refers to the $m = 41$; $n = 40$ transition in hydrogen, and He50γ refers to the $m = 53$; $n = 50$ transition in helium.

The transition probability is given by equation (2.5). When $n \gg 1$ and $m - n \ll n$, the oscillator strength can be approximated by (Brown, 1987):

$$f_{mn} = f_{(m-n)} \left(1 + \frac{3\,(m-n)}{2n} \right) \tag{6.26}$$

with $f_{(m-n)} = 0.19077, 0.026332, 0.0081056$ and 0.0034917 for the α, β, γ and δ transitions, respectively. From this equation and (2.5), the H$n\alpha$ transition probability is therefore:

$$A_{n\alpha} \sim 1.167 \times 10^9 / n^6 \tag{6.27}$$

and the radiative decay timescale, $\tau_{rad} = (A_{n\alpha} + A_{n\beta} + ..)^{-1}$. Since the orbiting electron is at large distance from the nucleus; $r = a_0 n^2$, where $a_0 = 0.529 \times 10^{-8}$ cm is the radius of the first Bohr orbit, we can take the electron to be localised in its orbital, and therefore we can assume that collisional redistribution of energy will take place on the free electron interaction timescale appropriate to a plasma with a temperature of order 10^4 K. This gives $\tau_{eq.} \sim 1.95 \times 10^4 / n_e$ s. Collisions with free electrons will produce LTE level populations in the excited bound states when $\tau_{eeq} \ll \tau_{rad}$. For hydrogen this condition yields:

$$n > 168 n_e^{-1/6} \tag{6.28}$$

as the principal quantum number above which which the transition to full LTE conditions begins. The transition will first occur amongst the different $l-$ levels, which have no energy barrier associated with them, but for collisional ionisations the energy barrier in effect means that the effective $\log \Lambda$ is smaller, the effective collisional timescale longer, and the transition to LTE will occur at higher n. Since the b_{nl} factors of the lower states are less than unity, db_n/dn is positive. These results are in fair agreement with detailed calculations (Seaton, 1964; Brocklehurst & Seaton 1972), and serve to illustrate the basic physics.

The line emission coefficient and absorption coefficient can be written in terms of the transition probabilities, and the Einstein coefficients (*c.f.* Section (4.1.3)). The emission coefficient is:

$$j_v^L = A_{mn} N_m h\nu \Phi(\nu), \tag{6.29}$$

where N_m is the population in the upper level, $\Phi(\nu)$ is a normalised line profile function such that $\int \Phi(\nu)d\nu = 1$. The absorption coefficient is:

$$\kappa_\nu^L = \left(N_n B_{nm} - N_m B_{mn} \right) h\nu \Phi(\nu), \tag{6.30}$$

or, in terms of the b_n factors from their definition in Section (5.4):

$$\kappa_\nu^L = N_n B_{nm} h\nu \Phi(\nu) \left(1 - \frac{b_m}{b_n} \exp\left[-h\nu/kT_e \right] \right). \tag{6.31}$$

Since $b_m > b_n$, and $h\nu/kT_e \ll 1$, then there is a stimulated emission effect to be taken into account in the line transfer and analysis of the radio recombination lines. The coupling of optical depth, and stimulated emission and line–continuum effects make the problem of the analysis of the recombination line observations a difficult and complex problem even for quite simple nebular geometries. Despite these problems, radio recombination lines can be used, and have been used to map the electron temperature distribution and to determine the ratio of ionised helium to ionised hydrogen in photoionised regions throughout the galaxy, with the result that N(H)/N(He)=0.09, with little evidence of variation according to the radial location in the galaxy.

Notes on Chapter 6

- The radio spectrum of ionised regions is fully discussed by Brown, R.L. 1987, in *Spectroscopy of Astrophysical Plasmas*, eds. A. Dargarno & D. Layzer, CUP:Cambridge, p. 35-58. (ISBN 0 512 23615 8).

Symbol		Definition
$_2F_1$	:	Standard hypergeometric function
α	:	Radiation power–law index in frequency units
a_0	:	The Bohr radius $\hbar^2/(m_e e^2)$, cm
A_{el}	:	The abundance of the element el by number relative to H
c	:	The speed of light, cm/s
$\chi_{el,Z}$	:	Ionisation fraction of element el in stage Z
χ_{ff}	:	Free–free emission as a fraction of the total emission Λ_{ff}
η_i, η_f	:	The scaled inverse initial and final electron energy, $\sqrt{Z^2 Ry/E}$
ε_i	:	The scaled initial electron energy, $E_i/Z^2 Ry$
$\mathbf{e}$	:	The electron charge in electrostatic units
E_i, E_f	:	The Initial and final electron energies, ergs
f_k	:	Free–free continuum constant for emission as a function of frequency
F_k	:	Free–free continuum constant for total emission
F_ν	:	Radiation flux in frequency units, ergs cm^{-2} s^{-1} Hz^{-1} sr^{-1}
Γ	:	The standard complex Γ function
γ^2	:	The scaled inverse temperature experienced by an ion, $Z^2 Ry/kT$
G_l	:	A matrix element, defined in terms of l, η_i, η_f and $_2F_1$
$h, \hbar$	:	Planck's constant and $h/2\pi$ respectively
I_l	:	An integral, defined in terms of $l, \eta_i, \eta_f, \Gamma$ and G_l
$I_{el,Z}$	:	Ionisation potential of element el in ion stage Z
k	:	Boltzmann constant, ergs K^{-1}
k_i, k_f	:	Wavenumbers of electrons in initial and final states
Λ_{ff}	:	Total free–free emission, ergs cm^{-1} s^{-1}
l	:	Angular momentum quantum number of electron state, 0, 1, 2 $\ldots \infty$
m_e	:	The electron mass, g
ν	:	Frequency of a photon, Hz
n_e, n_H	:	The number densities of electron and hydrogen atoms, respectively
Ψ_l	:	Electron wavefunction for state l
P_{ff}	:	Free–free emissivity as a function of ν and T
Q_ε	:	Ionisation parameter above energy ε, cm s^{-1}
Q_H	:	Ionisation parameter above 13.59844eV, cm s^{-1}
r	:	Radial coordinate
ρ_f, ρ_i	:	Quantum state densities
Ry	:	Rydberg energy, ergs , $2\pi^2 \mathbf{e}^4 m_e/h^2$
$\sigma_{i,\nu-f}$	:	Cross–section for free–free absorption of a photon ν by state i
$\sigma_{f-i,\nu}$	:	Cross–section for free–free photon emission of a photon ν by state f
s_1, s_2, s_3	:	Spline coefficients
T, T_e	:	Electron temperature, K
u	:	Thermal scaled photon energy, $h\nu/kT$
U	:	Dimensionless ionisation parameter, Q_ε/c
v_f	:	Velocity of electron in final state f
w	:	Scaled photon energy, $h\nu/Z^2 Ry$
x	:	Dummy variable
Y	:	Mass fraction of helium
z	:	A complex number
Z	:	Ion charge

7. Cooling Plasmas

"What after all is a halo? Its only one more thing to keep clean"
— Cristopher Fry (The Lady's not for burning, 1949)

7.1 The Cooling Function

The total energy radiated by a plasma per unit volume and per unit time is the sum of the various contributions to the local cooling (or heating). In general, it is a function of temperature, density, and chemical abundances of the elements:

$$\dot{Q}\,(n_e, T_e, Z_A) = \dot{Q}_{line} + \dot{Q}_{coll} - \dot{Q}_{rec} + \dot{Q}_{cont} - \dot{Q}_{phot} \pm \dot{Q}_{Compt.} \quad (7.1)$$

where the various heating or cooling terms represent collisional excitation cooling through emission lines ($\dot{Q}_{line}$), collisional ionisation losses ($\dot{Q}_{line}$), heating through recombinations ($\dot{Q}_{rec}$), emission through continuum process ($\dot{Q}_{cont}$), heating by photoionisation ($\dot{Q}_{phot}$), and Compton heating or cooling ($\dot{Q}_{Compt.}$). Two points should be noted here:

First, the recombination term is a heating effect locally, since recombinations preferentially remove the lower energy electrons from the electron gas, because it is for these electrons that the density of virtual states in the continuum is greatest. Often in the past, recombinations have been regarded as a loss process, in view of the recombination line emission that they produce. However, if we think about this for a moment, we will realise that the energy which these lines represent was removed earlier, at the time of the collisional ionisation of the species, since this is the moment when an energy equal to the ionisation potential of the ion was extracted from the electron gas. The fact that this energy is emitted later, after recombination, is in some sense immaterial, since the recombination line emission does not affect the local thermal balance of the plasma. This delicate distinction does not much matter for a plasma in photoionision equilibrium, since the rate of ionisations and the recombinations are in temporal balance in a given spatial region. However, it becomes very important in cooling plasmas where the regions of ionisation and of recombination may be quite separated in space and / or time.

Second, the *Compton heating* term accounts for the heating (or cooling) effect of inelastic or superelastic scattering of photons by the electron gas. This term may occur with either sign, depending on the ratio of the thermal energy of the electrons and the mean energy of the photons that are being scattered. From Krolik *et al.* (1981), the Compton term is given by:

$$\dot{Q}_{Compt.} = \int_\nu F_\nu \sigma_{v,\gamma} \frac{n_e}{m_e c^2} \left(4kT_e - h\nu \right) d\nu \tag{7.2}$$

where F_ν is the flux of radiation at frequency ν, and $\sigma_{v,\gamma}$ is the scattering cross - section which is dependent on the ratio of the energy of the photon, $h\nu$, and the energy of the electron. Expressing the electron energy as $E = \gamma m_e c^2$, the scattering cross - section is given by the (fully relativistic) Klein-Nishina formula:

$$\sigma_{v,\gamma} = \frac{3\sigma_T}{4} \left\{ \frac{(1+q)}{q^3} \left(\frac{2q(1+q)}{(1+2q)} - \ln[1+2q] \right) + \frac{1}{2q} \ln[1+2q] - \frac{(1+3q)}{(1+2q)^2} \right\}$$

with $q = \dfrac{\gamma h\nu}{m_e c^2}$ and $\sigma_T = (8\pi/3)\left(e^2/m_e c^2\right)^2 = 6.65 \times 10^{-25} \text{cm}^2$ \hfill (7.3)

Here, σ_T is simply the Thompson electron scattering cross - section. In the non-relativistic thermal plasmas that we are dealing with here $\langle \gamma \rangle = \frac{3kT_e}{2m_e c^2}$. This scattering cross - section takes the following asymptotic forms:

$$\sigma_{v,\gamma} = \sigma_T : q \ll 1$$
$$\sigma_{v,\gamma} = \frac{3\sigma_T}{8q} \left\{ \frac{1}{2} + \ln[2q] \right\} : q \gg 1 \tag{7.4}$$

Because the difference of the thermal and the photon energy appears in equation (7.2), the Compton term changes sign when $q = 8\gamma^2/3$.

In computations of optically thin, low density cooling plasmas, the photon density is usually assumed to be so low (and/or the plasma so highly ionised) that both the photoionisation and the Compton terms can be neglected. In addition, in the coronal approximation, only the ground state of any particular ion is populated to any appreciable extent. In this limit, all rates of the remaining processes depend on two-body interactions, and any term in the heat loss equation can be written as:

$$\dot{Q}\left(n_e, T_e, Z_A\right) = n_e n_A R^A(T_e) = n_e n \Lambda(T_e, Z_A) \tag{7.5}$$

where n is the total density of atoms and ions per unit volume. Writing the heat loss in this form allows the density dependence to be separated from the temperature dependence. The cooling function of the plasma (in units of erg cm^3s^{-1}) is therefore:

$$\Lambda(T_e, Z_A) = \Lambda_{line} + \Lambda_{coll} - \Lambda_{rec} + \Lambda_{cont} \tag{7.6}$$
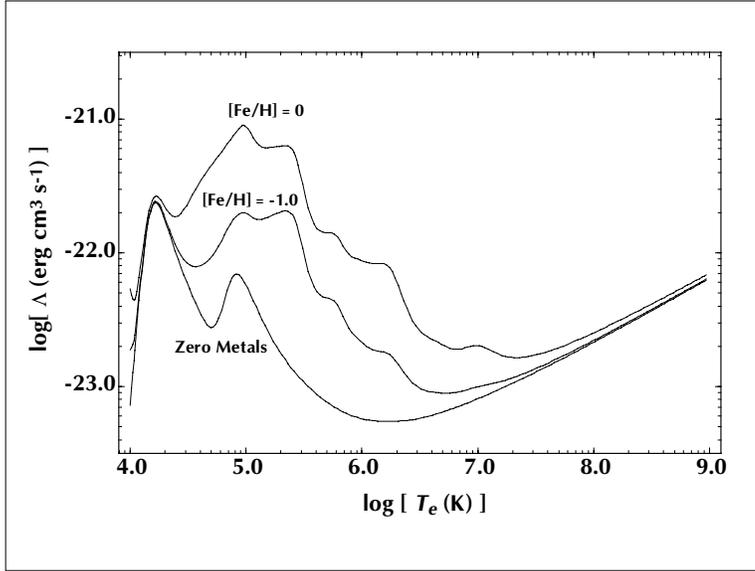
Fig. 7.1. The collisional-ionisation cooling function as a function of metallicity. At zero "metals" (elements heavier than helium) the peaks due to hydrogen and helium are apparent, while those of carbon, oxygen, neon, silicon and iron become apparent in the other curves. After Sutherland & Dopita (1993).

Figure (7.1) shows the cooling function derived for collisional ionisation equilibrium conditions as a function of the heavy element abundance, from Sutherland & Dopita (1993).

When only hydrogen and helium are present, the cooling function is dominated by the collisional excitation of excited states in H^0 and He^0 for the temperature range $\log[T_e] \lesssim 4.6$. For higher temperatures, up to $\log[T_e] \sim 5.6$, collisional excitation of He^+ dominates the cooling. At higher temperatures still, the electron Bremsstrahlung (free-free) cooling is dominant. The X-ray spectrum of such a Bremsstrahlung-dominated plasma is very simple, and is given by the sum of the free-free emissivity of all the ions present:

$$\epsilon_{ff}(T_e, Z_A, \nu) = \frac{2^5 \pi e^6}{3 m_e c^3} \left(\frac{2\pi}{3 m_e k}\right)^{1/2} \left(\sum_{A,i} g_{ff}(T_e, i, \nu) n_{A,i} i^2\right)$$

$$\text{x } n_e T_e^{-1/2} \exp\left[-\frac{h\nu}{kT}\right] \tag{7.7}$$

Thus νF_ν (which measures the spectral distribution in energy) increases linearly until $h\nu \sim kT$, after which it cuts off exponentially. This is shown in Figure (7.2). As the metal abundances approach solar, collisional excitation of the heavy elements provides the dominant cooling term, and continue to do
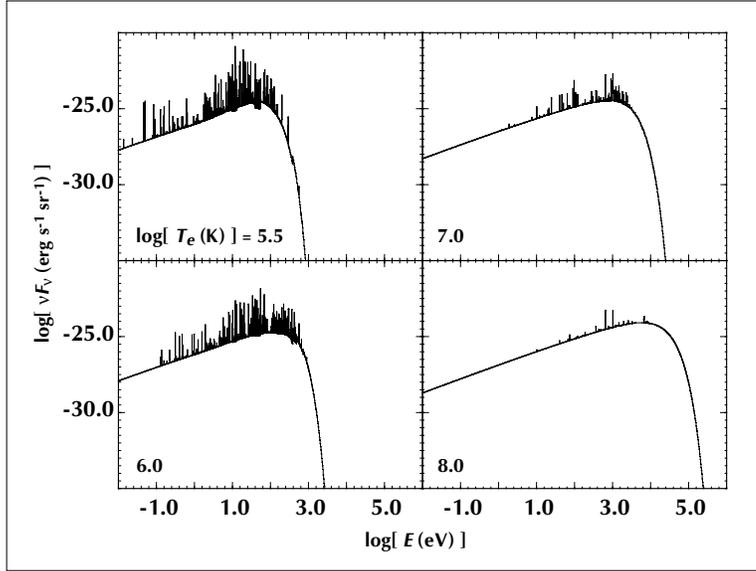
Fig. 7.2. The CIE X-ray spectra for different electron temperatures. The underlying shape of the electron Brehmsstrahlung spectrum becomes apparent at high temperatures, while at lower temperatures this is drowned by the emission lines, free-free and bound-free continuua of heavy elements. After Sutherland & Dopita, (1993).

so with increasing temperature until the K-shell electrons are finally stripped from Fe; above $\log[T_e] \sim 8.0$. In the temperature regime $\log[T_e] \gtrsim 7.3$ the Fe K line at 7 keV is the strongest in the spectrum, and provides a powerful X-ray diagnostic tool for such hot plasmas. Up to about 10^5 K the principal emission lines are found in the optical/UV region of the spectrum, but above this, they are located mainly in the EUV and X-ray regions of the electromagnetic spectrum.

7.2 Conditions for Non-Equilibrium Cooling

The total heat content of the plasma (energy per unit volume), assumed to be monoatomic and with an ion temperature T_{ion} which is the same as the electron temperature, is $Q = 3/2(n_e + n)kT_e$, where n is the total ion density. The assumption that $T_{ion} = T_e$ may break down in certain highly non-equilibrium situations, as we will see later. Combining this heat content with the heat loss equation (7.5) leads to the local cooling timescale:

$$\tau_{cool} = \frac{Q}{\dot{Q}} = \frac{3(n_e + n)kT_e}{2n^2 \Lambda(T_e, Z_A)} \tag{7.8}$$

Frequently, the cooling timescale is simplified as a simple function of temperature and the inverse of the density, by fitting the $\Lambda(T_e, Z_A)$ to a simple power law of the temperature, determined by the slope of the local tangent to the cooling function on a log : log scale; $\Lambda(T_e) = \Lambda_0 T_e^p$. In this case, assuming that all the electrons are supplied by either hydrogen or helium and that both are assumed to be fully ionised):

$$\tau_{cool} \sim \frac{3k\,[1 + 2.5Z(He)]}{\Lambda_0\,[1 + Z(He)]} T_e^{1-p} n^{-1} \tag{7.9}$$

where $Z(He) = n_{He}/n_H \sim 0.1$ in most diffuse astrophysical plasmas.

This approximation is exact at the high temperatures where the electron bremsstrahlung is the dominant cooling term and the cooling function is then $\Lambda \sim 3 \times 10^{-27} T_e^{1/2}$. In this case $\tau_{cool} \sim 5 \times 10^7 n T_8^{1/2}$ years, where n is the total atom density (cm^{-3}) and T_8 is the temperature in units of 10^8 K. In the Universe, the longest time available is of order of the Hubble timescale, $\tau_{Hubb} \sim 1.5 \times 10^{10}$ years. This implies that plasmas with densities $\leq 3 \times 10^{-3}$ cm^{-3} which have been heated to a temperature of 10^8 K or greater can never cool again.

The idea that collisional ionisation equilibrium represents a realistic state for the plasma is founded on two assumptions. The first is an implicit assumption; the so-called coronal approximation discussed in section (5.5). This takes the plasma to be optically thin to the escape of all radiation produced within its volume, ignores photoionisation, and assumes that only the population in the ground term of each ionic species is populated to any appreciable degree. The second requires that equilibrium between the various collisional processes has been achieved. This may not necessarily be the case, even if the conditions of the coronal approximation are valid. Assume the plasma is at a sufficiently high temperature that the charge-exchange reactions can be neglected, and consider the time-dependent ionisation balance equation for any ion of any atom; A^{i+}:

$$\frac{dn_{A,i}}{dt} = n_e n_{A,i+1} \alpha_{rec}^{A,i+1} - n_e n_{A,i} \alpha_{rec}^{A,i} - n_e n_{A,i} R_{coll}^{A,i} + n_e n_{A,i-1} R_{coll}^{A,i-1} \tag{7.10}$$

or, introducing the fractional ionisation $x_{A,i} = n_{A,i}/n_A$ this is written:

$$\frac{1}{x_{A,i}} \frac{dx_{A,i}}{dt} = n_e \left(\frac{x_{A,i+1}}{x_{A,i}} \alpha_{rec}^{A,i+1} - \alpha_{rec}^{A,i} - R_{coll}^{A,i} + \frac{x_{A,i-1}}{x_{A,i}} R_{coll}^{A,i-1} \right) \tag{7.11}$$

When the plasma is under-ionised as compared with its equilibrium value, then it is driven towards the equilibrium value by collisional ionisations from lower ionisation states, so that the second and third terms in this equation dominate over the first. When the gas is over-ionised for its temperature, then the recombination term from the more highly ionised states is the key factor in determinining the approach to equilibrium. Thus, we can define the *collisional equilibration timescale*, τ_{CIE}, as:
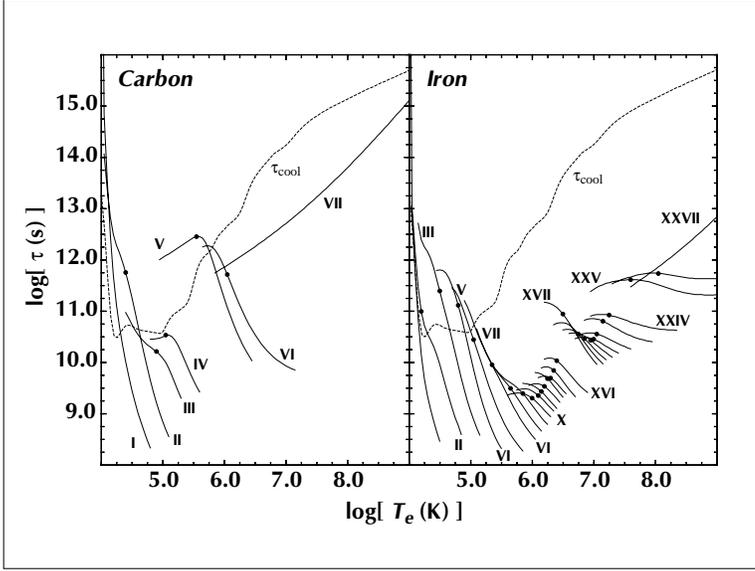
Fig. 7.3. The temperature dependence of the collisional ionisation equilibration timescale for the various ions of carbon (left) and of iron (right) are shown here compared with the cooling timescale for a plasma of solar composition. The assumption of CIE clearly breaks down at low temperatures, low ionisation state and for the lighter elements.

$$\tau_{CIE} = \left(\frac{x_{A,i+1}}{x_{A,i}} \alpha_{rec}^{A,i+1} - \alpha_{rec}^{A,i} - R_{coll}^{A,i} + \frac{x_{A,i-1}}{x_{A,i}} R_{coll}^{A,i-1} \right)^{-1} n_e^{-1} \qquad (7.12)$$

For most purposes, the approximation:

$$\tau_{CIE} = \frac{1}{\left(\alpha_{rec}^{A,i} + R_{coll}^{A,i} \right) n_e} \qquad (7.13)$$

provides an adequate means of estimating the equilibration timescale.

Because of this finite collisional equilibration timescale, a plasma which cools too quickly will never be able to come into equilibrium at its instantaneous temperature. Such rapid cooling is clearly favoured when the abundances of the heavy elements in the plasma are high, since it is these elements which provide most of the cooling. The condition that collisional ionisation equilibrium is valid in a coronal plasma is simply:

$$\tau_{cool} \gg \tau_{CIE} \qquad (7.14)$$

for all the ionic species present.

Does this condition apply in the case of a plasma with solar abundances? In Figure (7.3) we show the cooling timescale for the plasma as a whole, and

the CIE timescales for the various ions of C and of Fe. For Fe, CIE is an excellent approximation for $\log T_e \gtrsim 5.5$, but it fails at lower temperatures than this thanks to the decrease in the cooling timescale, and the increasing recombination timescale for species of low ionic charge. For C, CIE is a fair approximation only for $\log T_e \gtrsim 6$, and clearly fails for the hydrogen-like and helium-like species which have particularly long recombination timescales. In general, this is true for all the lighter atomic species.

The departures from CIE become even more apparent for a plasma that has been heated suddenly from a much lower temperature. Since, initially, the plasma is very much under-ionised compared with its equilibrium value, both Λ_{line} and Λ_{coll} are strongly enhanced over their equilibrium values. This has the consequence that τ_{cool} is much smaller than in the equilibrium plasma of the same temperature. Line cooling from collisional excitation of helium- and hydrogen- like Lyman lines is particularly important, since the ionisation timescale of such ions is long, and many collisional excitations occur for each ionisation. In such plasmas, the initial temperature must be such that $\log T_e \gtrsim 7.5$ for the plasma to "forget" its initial ionisation state, and settle down into CIE during the subsequent cooling.

In a plasma which is undergoing non-CIE cooling, the plasma eventually reaches a state in which the recombination timescales become long compared with the cooling timescales and the plasma ionisation lags the equilibrium at that temperature. As a result, it is over-ionised. In this state Λ_{line} is surpressed compared with its equilibrium value, because the temperature is too low for the electrons to overcome the larger energy barriers for collisional excitation of the more highly ionised ions which are present. As a consequence the whole shape of the cooling function is altered, as shown in Figure (7.4).

In summary, collisional ionisation equilibrium is an approximation which is fully valid only in very hot plasmas. At low temperatures, all cooling plasmas are strongly out of CIE. The temperature at which the CIE approximation fails depends strongly on the relative importance of the line cooling, which is a strong function of both the abundance of the heavier atoms and of the thermal history of the plasma, and which dominates the cooling function below $\log T_e \gtrsim 7.0$.

7.3 Heat Transport

Advanced Topic

The non-CIE effects in a cooling plasma which were described in the previous section may to some extent be vitiated if the plasma is able to redistribute the thermal energy content from the hotter to the cooler regions. This can occur as a result of heat conduction or through wave or turbulent transport. We will consider here the extent to which these may be important.
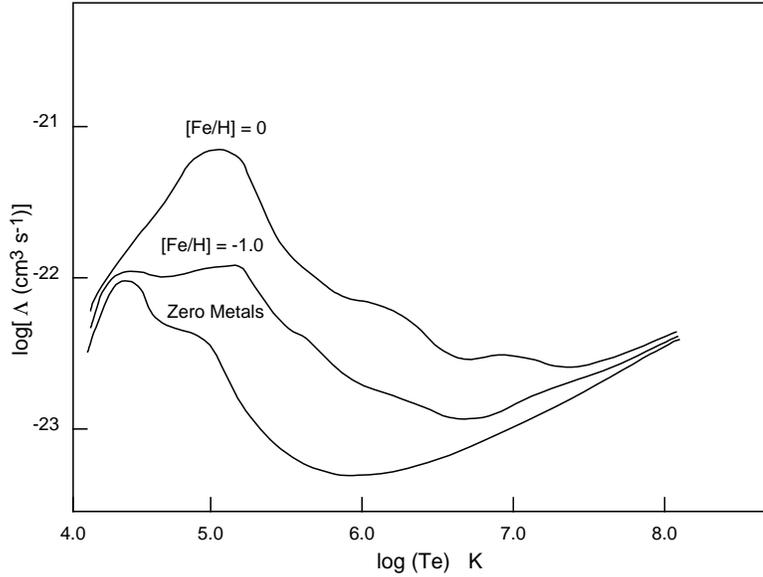
Fig. 7.4. The cooling function under conditions of non-equilibrium cooling as a function of metallicity. Note how the peaks of Figure (7.1) are smeared out. After Sutherland & Dopita, 1993

7.3.1 Electron Conduction

In a thermally conductive medium, the heat flux carried by *electron conduction* per unit area follows from the definition of the *thermal conductivity*, κ, by $\dot{\mathbf{q}} = -\kappa \nabla \mathbf{T}$. In a plasma with a shallow temperature gradient, each electron suffers many collisions over the scale length of the temperature gradient. In this case the conductivity is directly proportional to the mean path for electron energy exchange, λ, which from Cowie & McKee (1977) is given by:

$$\kappa = 1.31 n_e k \lambda \left(\frac{kT_e}{m_e} \right)^{1/2} \tag{7.15}$$

Here, the numerical factor takes into account the fact that thermal transport by electrons cannot cause an electric field to be set up. This reduces the conductivity by a factor of 2.5 (under most conditions) from the factor that kinetic theory might otherwise have suggested (Spitzer, 1956). The *mean free path* of the electrons is expressed in terms of the *electron equipartition timescale*, τ_{eeq} and the mean electron thermal velocity, $v_e = (3kT_e/m_e)^{1/2}$:

$$\lambda = \tau_{eeq} \left(\frac{3kT_e}{m_e} \right)^{1/2} \tag{7.16}$$

so that to compute the electron conductivity, it remains to estimate τ_{eeq}. This timescale is determined by the timescale over which each electron loses all information about its initial drift velocity relative to the bulk of the electron gas as a result of collisions with other electrons. In a collision between two electrons, each is scattered through an angle $(\pi - 2\psi)$ where $\tan\psi = bm_e v_e^2/2e^2$ where b is the impact parameter, the distance between the centres of the two approaching particles measured along the trajectories at infinity. A collision which is effective in redistributing the momentum of the colliding particles is one in which the electron paths are deflected by 90°, *i.e.* $\tan\psi = 1$, in which case $b = b_0$. We can therefore write the *electron-electron collision timescale* τ_{ecoll} as:

$$\tau_{ecoll} = \left(\pi b_0^2 n_e 2v_e\right)^{-1} = \frac{3^{3/2} m_e^{1/2}}{2\pi n_e e^4}\left(kT_e\right)^{3/2} \tag{7.17}$$

τ_{eeq} is shorter than this τ_{ecoll} because of the combined effects of many distant collisions, which alter the momentum vector of the colliding electrons only a little. The fractional change in momentum when ψ is in the small angle limit is given by $dp/p = 2bm_e v_e^2/e^2$ and the number of collisions increases as πb^2. As a result, the equipartition timescale is shorter than the collision timescale defined above by a (Coulomb integral) factor of approximately

$$\frac{\tau_{ecoll}}{\tau_{eeq}} = 2\int_{b_{\min}}^{b_{\max}} b^{-1}db = 2\ln\left(\frac{b_{\max}}{b_{\min}}\right) = 2\ln\Lambda_{coul} \tag{7.18}$$

where the upper limit to the integral is determined by Coulomb shielding effects of the other charged particles present. An exact calculation (Spitzer, 1956) gives:

$$\tau_{eeq} = \frac{3m_e^{1/2}\left(kT_e\right)^{3/2}}{4\pi^{1/2}n_e e^4 \ln\Lambda_{coul}} \tag{7.19}$$

with

$$\Lambda_{coul} = \frac{3\left(kT_e\right)^{3/2}}{2\pi^{1/2}n_e^{1/2}e^3}\left[1 + \frac{v_e}{2\alpha c}\right]^{-1} \tag{7.20}$$

Here the second term in the square brackets allows for the effect of electron diffraction which modifies the classical result when the thermal velocity of the electrons divided by the speed of light becomes comparable with the fine structure constant $\alpha = e^2/hc = 1/137.1..$. Putting all these results together, we obtain the Classical formula given by Spitzer (1962) for the thermal conductivity of a fully ionised plasma:

$$\kappa = 1.84 \times 10^{-5} T_e^{5/2}/\ln\Lambda_{coul}\ \text{ergs}^{-1}\text{cm}^{-1}\text{K}^{-1} \tag{7.21}$$

where for $T_e > 4 \times 10^5$ K;

$$\ln\Lambda_{coul} = 32 + \ln\left[n_e^{-1/2}\left(T_e/10^7 K\right)\right] \tag{7.22}$$

When a cool cloud is in contact with a coronal medium, there will be a very steep temperature gradient at the boundary. In this circumstance equation (7.21) may break down, since the basic assumption involved in its derivation, that many collisions take place over the scale length of the temperature variation, is invalid. As a result, equation (7.21) grossly overestimates the importance of electron conduction. Cowie and McKee (1977) suggested that in the limit of steep temperature gradients, the heat flux is limited simply by the ability of the electrons to diffuse across the boundary between the hot and the cold medium. The limiting heat flux through the boundary is therefore:

$$\dot{q} = \int_0^{\pi/2} \frac{m_e n_e v_e^3}{2} \cos\theta d\theta = \frac{3}{2} \left(\frac{8}{9\pi}\right)^{1/2} \left(\frac{kT_e}{m_e}\right)^{1/2} n_e kT_e$$

$$= \left(\frac{2}{\pi}\right)^{1/2} \left(\frac{kT_e}{m_e}\right)^{1/2} n_e kT_e \qquad (7.23)$$

Cowie & McKee further assume that this flux is reduced by 2.5, the same factor as for classical conductivity, as a result of the requirement that the conduction be electric field-free. With this modification, equation (7.23) provides an estimate of the *saturated* electron heat conductivity.

Any magnetic field which is present has a dramatic effect on the thermal conductivity, since the *gyroradius* of thermal electrons about the field lines, $r_{eg} = (m_e v_\perp c/eB) = (2m_e kT_e)^{1/2} c/eB$, is quite small compared with any other significant scale length.

In the presence of a magnetic field the conductive heat flux is reduced to:

$$\dot{\mathbf{q}} = -\kappa \left(\frac{\mathbf{B}}{B}\right) \wedge \nabla \mathbf{T} \qquad (7.24)$$

This is lower than the field-free case by a factor $\cos\varphi$ where φ is the angle between the local field direction and the direction of the temperature gradient. Since conduction along field lines is not inhibited, while conduction across field lines is supressed, the regions of cooled plasma will tend to form thin filaments in the direction of the field lines. Because the electrical conductivity of a fully ionised plasma is very high, the field is pinned to gas as it cools, and the magnetic field increases as the gas cools isobarically. Eventually the pressure due to the transverse magnetic field will match the gas pressure in the hot medium, $B^2/8\pi = (n + n_e) kT_e$, after which further compression in this direction becomes impossible.

7.3.2 Boundary Layer Mixing

A second means of transporting energy from the hot to the cool meduim is through *boundary layer mixing*. This will occur most efficiently in astrophysical plasmas if there is a velocity shear across the boundary layer. Such a

flow could exist in many astrophysical situations. Examples that could be cited are in during the breakout of a hot bubble of gas produced by winds or supernova explosions from the plane of the galaxy, in the outflow of hot gas from active galactic nuclei, or as a result of *Rayleigh-Taylor* instabilities in the hot ejecta of supernovae (Rayleigh-Taylor instabilities occur when a low density medium tries to accelerate a denser medium of the same pressure. The boundary layer breaks up into dense blobs, which are left behind expanding bubbles of the low density fluid).

Shear flows between two fluids of different temperatures but the same pressure excite the *Kelvin-Helmholz* instabilities which initially appear as thransverse travelling waves along the boundary between the two fluids, which grow, become breakers, so that a layer of turbulent mixing is formed. The mixing cascades in turbulent eddies down to the dissipative level, which produces a layer of gas at an intermediate temperature. In this layer the energy flux input by turbulent motions is balanced by the cooling rate per unit area, for as long as the hot gas remains to supply the heat flux. Such mixing layers were proposed by Begelman & Fabian (1990) and the theory was developed by Slavin, Shull and Begelman (1993). If T_H is the temperature of the hot medium, and T_C the temperature of the cold, and v_{turb} is the mean turbulent velocity in the mixing layer, then the mean temperature of the mixed layer is given by:

$$\langle T_e \rangle = \frac{\dot{m}_H \, T_H + \dot{m}_C \, T_c}{\dot{m}_H + \dot{m}_C} \tag{7.25}$$

with

$$\dot{m}_H = \eta_H \rho_H v_{turb}$$
$$\dot{m}_C = \eta_C \left(\rho_H \rho_C \right)^{1/2} v_{turb} \tag{7.26}$$

and the heat flux balance condition in the layer is given by:

$$\dot{q} = \Lambda \langle T_e \rangle \, n n_e \, \langle s \rangle = \frac{5}{2} \left(n + n_e \right) k \, \langle T_e \rangle \, \eta_H v_{turb} \tag{7.27}$$

where $\langle s \rangle$ is the thickness of the boundary layer. The turbulent velocity is somewhat less than, but of the same order as the sound speed of the gas at the temperature $\langle T_e \rangle$. Typically this will be a few tens of km s^{-1}. The main uncertainty in these equations is the size of η_H, the fraction of the mass in the mixing layer contributed by the hot medium, and of η_C, the efficiency of the hydrodynamical mixing, which cannot be estimated *a priori*, and which require a detailed high-resolution hydrodynamic model to calculate.

7.4 Cold Clouds in Hot Gas

In principle, it is possible to find an equilibrium in which small cold clouds exist in hot gas. In this case, the pressure in the cloud and in the surrounding

hot medium must be the same so that $\rho_C/\rho_H = T_H/T_C$. However, conduction and cooling can disrupt this balance. If more heat is conducted into the cold cloud than can be radiated away, then the cloud must evaporate. If the reverse is true, then cooling will dominate around its boundary, and instead it will accrete matter from the hot medium. A given parcel of gas will be heated by conduction only if the conductive energy flux exceeds the radiative losses within it. The ratio of conduction to radiation is given by (Begelman & McKee, 1990):

$$\frac{\nabla . \kappa \nabla T}{\Lambda n^2} \sim \frac{\kappa T}{L^2} . \frac{1}{\Lambda n^2} \tag{7.28}$$

where L is a distance about equal to the temperature scale height. We may therefore define a *Field Length*, λ_F, equal to:

$$\lambda_F = \left(\frac{\kappa T}{\Lambda n^2} \right)^{1/2} \tag{7.29}$$

From equation (7.28), it is clear that when the scale length is less than λ_F thermal conduction will dominate over radiative processes, and when the scale length is greater than λ_F radiative losses are dominant. The Field length therefore determines the fate of small clouds embedded in a hot medium (neglecting any effects of turbulent transport of heat). If the cloud radius is less than λ_F the cloud will rapidly evaporate, while if it is larger than this, it will be stabilised by radiative losses around its periphery.

Consider a spherical cloud, and assume that there is a steady, constant pressure evaporative flow with mass flux $\dot{M}$ driven from it by the conductive heating. Matching the heat transport through a surface gives the equation of motion (McCray, 1987):

$$\frac{5}{2} \frac{\dot{M}}{\mu m_H} \frac{k}{dr} \frac{dT_e}{dr} = \frac{d}{dr} \left(4\pi r^2 \kappa \frac{dT_e}{dr} \right) \tag{7.30}$$

where μ is the mean atomic weight. In the case of classical conductivity $\kappa = \zeta T_e^{5/2}$; ζ being a constant, which from equations (7.21) and (7.22) is about 5.7×10^{-7}. Equation (7.30) has the solution:

$$T_e(r) = T_H \left(1 - R_C/r \right)^{2/5} \tag{7.31}$$

for evaporating clouds with cloud radius $R_C < \lambda_F$, the mass flux is:

$$\dot{M} = \frac{16\pi \mu m_H \zeta}{25k} R_C T_H^{5/2} \sim 1.1 \times 10^{-4} R_1 T_7^{5/2} M_\odot \text{yr}^{-1} \tag{7.32}$$

where R_1 is the cloud radius in units of parsecs (3.0856×10^{18} cm) and T_7 is the intercloud temperature in units of 10^7 K. Taking the pressure equilibrium condition between the cloud and intercloud medium, this implies an evaporative destruction timescale for the cloud:

$$\tau_{evap} = \frac{25k}{12\mu m_H \zeta} \left(\frac{T_H}{T_C}\right) \rho_H T_H^{-5/2} R_C^2$$

$$\sim 8.5 \times 10^5 \left(\frac{T_H}{T_C}\right) n_{-2} T_7^{-5/2} R_1^2 \text{ yr.} \tag{7.33}$$

where n_{-2} is the number density in the hot medium in units of 10^{-2} cm^{-3}. If we assume that the cold medium is in the molecular form, then we can expect $T_C \sim 100$K, so for $T_7 \sim 1, T_H/T_C \sim 10^5$. The cold cloud will survive if the hot medium cools down within the evaporative lifetime; $\tau_{evap} > \tau_{cool}$. The cooling time for the hot medium will be less than or about equal to the free-free cooling timescale we estimated following equation (7.9); above so clouds will certainly survive if $\tau_{evap} > \tau_{ff}$. Since $\tau_{ff} \sim 1.6 \times 10^9 n_{-2}^{-1} T_7^{1/2}$, using the same units as those of equation (7.33), the condition for cloud survival becomes:

$$R_C > 0.14 T_7^{3/2} \text{ pc.} \tag{7.34}$$

However, this size is much smaller than the electron-electron collision length estimated from equations (7.16) and (7.19); $\lambda \sim 32 T_7^2 n_{-2}^{-1}$ pc, so small clouds will be in the saturated electron conduction regime. We leave it as an exercise for the reader to estimate the survival condition of such clouds.

At the critical radius where the conduction into the cloud matches the radiative losses in the cloud, there is no bulk flow of material either into, or away from, the cloud. In this case the equation of energy balance becomes:

$$\frac{d}{dr}\left(\kappa \frac{dT_e}{dr}\right) + \frac{2}{r}\kappa \frac{dT_e}{dr} - n_e n \Lambda(T_e) = 0 \tag{7.35}$$

McKee and Cowie (1977) have solved this equation in a semi-analytic fashion by dividing the solution into zones where either the radiative losses are greater than the spherical divergence of the heat flux, or where the reverse is true, and by fitting the cooling function to power-law segments, as we described in section (7.2). Their results on the critical radius can be approximated by:

$$R_{crit} \sim 2000 T_7^2 n_{-2}^{-1} \text{ pc.} \tag{7.36}$$

which implies that, at these pressures ($P/k \sim 10^5$ K cm^{-3}) subsonic accretion by cooling is unimportant.

In summary, very small clouds can be destroyed by thermal conduction in the saturated regime, while the larger clouds may initially lose only a thin outer layer before the hot gas cools.

7.5 Thermal Instabilities

The problem of thermal instabilities was properly addressed for the first time in a classical paper by Field (1965), and here we follow the exposition

given by Balbus (1995). The basic idea is simple. A region with a certain temperature will cool on a timescale $\tau_{cool}(n, T_e, Z_A)$ given by equation (7.8). Now, consider an embedded region with a slightly lower temperature, and assume that any expansion or contraction is homogeneous. The medium will be thermally unstable if the temperature contrast between the warmer and the cooler regions increases with time. This requires that the cooling timescale in the cooler region be less than the cooling timescale in the hotter region. If the cooler region is sufficiently small that sound waves can travel across it, then the pressures in the hotter and the cooler region will equalise during the cooling, and the cooling (apart from any ram-pressure terms) will occur *isobarically*. On the other hand, if the region is large, the cooling will (at least initially) be at constant density. This is called *isochoric* cooling. Because the cooling timescale goes inversely as the square of the density (equation 7.8) and the density increases as T_e^{-1}, in isobaric cooling the cooling timescale in the lower temperature region decreases more rapidly during the cooling, which tends to make the instability even worse.

7.5.1 In a Stationary Medium

The second law of thermodynamics states that:

$$T_e \frac{dS}{dt} = -(\gamma - 1)\, \dot{Q} \tag{7.37}$$

where S is the entropy per gram, and $\gamma = 5/3$ for a monatomic gas. Now the gas is thermally stable if, in a perturbation of the entropy of a fluid element from S to $S+\Delta S$, the magnitude of ΔS tends to decrease with time. However,

$$\frac{d\ln|\Delta S|}{dt} = -\frac{2}{3}\left[\frac{\partial\left(\dot{Q}/T_e\right)}{\partial S}\right]_A \tag{7.38}$$

where A is the thermodynamic variable (pressure for isobaric disturbances or density for isochoric perturbations) held constant. The medium is therefore thermally stable if:

$$\left[\frac{\partial\left(\dot{Q}/T_e\right)}{\partial S}\right]_A > 0 \tag{7.39}$$

Now, the change in entropy for a process involving an ideal gas with gas constant $\Re$, specific heat at constant volume, C_V, and specific heat at constant pressure, C_p, is:

$$\begin{aligned} dS &= C_P \frac{dT}{T} - n\Re\frac{dP}{P} \\ &= C_V \frac{dT}{T} + n\Re\frac{dV}{V} \end{aligned} \tag{7.40}$$

So that the stability condition, equation (7.39) reduces to the Parker (1953) criterion:

$$\left[\frac{\partial \dot{Q}}{\partial T_e}\right]_\rho > 0 \tag{7.41}$$

for isochoric perturbations and to the Field (1965) criterion:

$$\left[\frac{\partial \dot{Q}}{\partial T_e}\right]_P > 0 \tag{7.42}$$

for isobaric perturbations. If the cooling is represented by a local power-law on the temperature, then $\dot{Q} = \Lambda_0 T_e^p n^2$. It follows that the medium is thermally stable in isochoric cooling if $p > 0$ and that it is thermally stable in isobaric cooling if $p > +1$. However, from the form of the non-CIE cooling function given in Figure (7.4), it is clear that p is rarely as large as unity, and is greater than zero only below (roughly) 10^5 K and above 10^7 K. Coronal plasmas are therefore usually thermally unstable, and will tend to break up into blobs or filaments.

Regions that are cooling isochorically are losing their internal pressure. If sound waves can carry from the hotter region into the cooler one on a timescale which is short compared with the cooling timescale, then cooling-induced pressure fluctuations will tend to be washed out, and the cooling is isobaric, as mentioned above. The *sound speed* in the ionised plasma, which we will term c_{II} (to distinguish it from sound speeds in un-ionised gas, c_I), is, for a perfect gas:

$$c_{II} = \left(\frac{dP}{d\rho}\right)^{1/2} = \left(\frac{\gamma P}{\rho}\right)^{1/2} = \left(\frac{\gamma k T_e}{\mu m_H}\right)^{1/2} \tag{7.43}$$

In a medium with a strong enough magnetic field, magnetic waves rather than sound waves may carry the information from one region to another instead. The *Alfvén velocity* $\mathbf{v}_A$, is defined as:

$$\mathbf{v}_A = \frac{\mathbf{B}}{(4\pi\rho)^{1/2}} \tag{7.44}$$

When the wave vector is aligned with the magnetic field, the magnetic *Alfvén waves* are transverse oscillations of the magnetic field, like waves in a skipping rope, and so do not help to equalise the pressure. In this case, as in the magnetic field-free case the pressure adjustment (a compression wave) is carried by the longitudinal sound waves. However, when the direction of the wave vector is perpendicular to the magnetic field direction, information is carried by longitudinal *magnetosonic* waves at the magnetosonic sound speed:

$$v_\perp = \left(c_{II}^2 + v_A^2\right)^{1/2} \tag{7.45}$$

Depending on which case applies, the condition for isobaric cooling is therefore:

$$R < c_{II}\tau_{cool} : \text{or } R < v_{\perp}\tau_{cool} \tag{7.46}$$

Since the sound speed and the cooling timescale both decrease with temperature, the size of the regions within which isobaric cooling is occurring becomes smaller and smaller, so that the medium tends to break up into a fractal heirarchy of blobs and filaments.

In practice, the rapid cooling of some parts of the plasma may cause their internal pressure to become decoupled from the surrounding hot medium. In this case, a compression wave from the hot medium steepens into a shock which serves to reheat and re-pressurise the cloud. The physics of such shock-waves is given in the next chapter.

7.5.2 In an Expanding Medium

Advanced Topic

In many fields of astronomy we have to deal with expanding clouds of hot gas. Examples include, the very early Universe before the epoch of recombination, the early phases of supernova explosions, and nova ejections, and in hot stellar winds. Here we will consider only plasmas in free expansion, so that the density of the substrate hot medium, ρ_H, behaves as $\rho_H = \rho_0 \left(t/t_o\right)^{-3}$. If this substrate had initially been heated up to a very high temperature then cooling is initially unimportant. However, it may well become important at a later phase due to the temperature decrease caused by the work done in the adiabatic expansion of the gas, $T_H = T_0 \left(t/t_o\right)^{-(\gamma-1)}$, where γ is the usual ratio of the specific heats (equal to 5/3 for a monatomic plasma).

We can distinguish between two types of growth of the thermal instability; *absolute growth*, in which both the density in the thermally unstable condensation and the density contrast between the blob and its surroundings increases in an absolute sense, and *relative growth*, in which the density contrast between the thermally unstable condensation and the surrounding medium increases but during which the density in the condensation is decreasing in an absolute sense. For absolute growth we have:

$$\frac{d\rho_C}{dt} > 0; \frac{d\left(\rho_C/\rho_H\right)}{dt} > 0 \tag{7.47}$$

here the second condition is redundant to the definition. According to the mathematical identity (the chain rule):

$$\frac{d\left(\rho_C/\rho_H\right)}{dt} = \frac{1}{\rho_H}\frac{d\rho_C}{dt} - \frac{\rho_C}{\rho_H^2}\frac{d\rho_H}{dt} \tag{7.48}$$

This implies that, since $\rho_C > 0, \rho_H > 0$ and $d\rho_H/dt < 0$, the second condition of equation(7.49) is automatically true in general. Now if we define two timescales:

$$\tau_C = \frac{\rho_C}{(d\rho_C/dt)} : \tau_H = \frac{\rho_H}{|d\rho_H/dt|} \tag{7.49}$$

then the condition for absolute growth can be simply written:

$$\frac{\tau_H}{\tau_C} > 0 \tag{7.50}$$

since ρ_C, ρ_H and $|d\rho_H/dt|$ are all greater than zero.

Now consider the conditions for relative growth:

$$\frac{d\rho_C}{dt} < 0; \frac{d\left(\rho_C/\rho_H\right)}{dt} > 0 \tag{7.51}$$

these are equivalent, using the identity (7.48) to;

$$0 > \frac{d\rho_C}{dt} > \frac{\rho_C}{\rho_H}\frac{d\rho_H}{dt} \tag{7.52}$$

or, in terms of the timescales defined above:

$$0 > \frac{\tau_H}{\tau_C} > -1 \tag{7.53}$$

For a uniform expansion, $\rho_H = \rho_0 \left(t/t_o\right)^{-3}$ and τ_H is simply related to the *Dynamical timescale* of the expansion, t_o; $\tau_H = t_o/3$. The timescale τ_C expresses the competition between the dynamical timescale which reduces the densities in the cloud, and the cooling timescale, which works to increase the cloud density; $\tau_C^{-1} = \tau_{cool}^{-1} - \tau_H^{-1}$, approximately. Thus, from equation (7.50) absolute growth will occur when the cooling timescale becomes shorter than one third of the dynamical timescale, and relative growth will start to occur from the beginning of the expansion, but will only become significant once the cooling timescale becomes of the same order as the expansion timescale. This suggests that it is valid to think of thermal instabilities in an expanding medium as occurring at a particular epoch in the expansion at which time blobs and filaments of cold material condense out of a previously more-or-less uniform substrate.

7.6 Hot Galactic Coronae

7.6.1 Early-Type Galaxies

X-ray observations with the *Einstein* and the ROSAT satellites have revealed that hot, extended galactic coronae are a common (but not universally observed) property of both spiral and early-type galaxies. For the E and S0 galaxies, the range of the X-ray observed luminosities is $43 > \log L_X > 39$, where the lower limit is determined effectively by the sensitivity limit of the

samples. There is a correlation with the optical luminosity, but this correlation shows a very large scatter of up to 100 in the X-ray luminosity for a given stellar luminosity. For the intrinsically more luminous galaxies, the X-ray luminosity rises more quickly than the stellar luminosity at optical blue wavelengths; $L_X \sim L_B^{1.7-2}$. For the X-ray faint galaxies, the X-rays appear to originate from stellar X-ray sources (binaries, catacylismic variables and the like), rather than from hot interstellar gas, which can be distinguished by its "softer" X-ray spectrum (Fabbiano, Kim & Trinchieri, 1992). For the more X-ray luminous galaxies, what then is the source of the hot, extended gas?

In the models of Ciotti *et al.* (1992), this comes from stellar mass-loss processes, either in the form of winds, or through the more violent thermonuclear (Type I) supernova explosions. Most of the stellar wind mass-loss occurs in the late stages of stellar evolution along the asymptotic giant branch, leading finally to ejection of planetary nebula (PNe) shells. These two processes eject material of quite different chemical composition. The stellar mass-loss is mostly from outer stellar envelopes which have been chemically unaltered since the star first formed out of the interstellar medium. However added to this are some products of incomplete hydrogen burning (mostly He, C and N and the heavy elements formed by slow neutron capture; the *s-process elements*). Type I supernova explosions, on the other hand, burn a fair fraction of the mass to nuclear statistical equilibrium composition at very high temperatures, leading to the ejection of mostly Fe and the iron-peak group of elements and heavy elements formed from rapid neutron capture in the explosion; the *r-process elements*. Thus, we may be certain that the composition of the X-ray haloes of early-type galaxies is unlike that of the local interstellar medium in our Galaxy. Rather, it must contain a greater fraction of heavy elements, which will lead to both an increase in the X-ray emissivity and a corresponding reduction in the cooling timescale of the gas.

Even when ejected in a hot phase from a stellar explosion, the the gas must cool again due to adiabatic losses. In order for it to be seen again as an diffuse X-ray emitter, it must be reheated by collisions in the galaxy. For the PNe, the source of this heating would be the energy associated with the orbital motion of the precursor star. For a stellar population of a given velocity dispersion, conventionally written as σ, collisions will heat the gas up to a temperature given by;

$$\frac{3}{2}kT_e \sim \frac{1}{2}\mu m_H \sigma^2 \qquad (7.54)$$

From the definition of the sound speed (equation 7.43), this implies that the effective Mach number of the stellar population with respect to the hot gas is $\mathcal{M} \sim 3/5^{1/2}$, or about 1.34. In this way, the gas temperature is related directly to the galaxian properties. For a typical galaxian velocity dispersion of 300 km.s^{-1}, the gas will reach a temperature of order 2×10^6 K ($kT_e \sim 150$ eV). In the case of Type I supernova ejecta the relevant velocity is the mean velocity of ejection of the material (~ 3000 km s^{-1}), which is much larger

than the velocity dispersion in the galaxy, and so this gas is initially heated to 2×10^8K ($kT_e \sim 15$ keV). The Type I ejecta mixes with the PN ejecta, so that the final temperature achieved by the gas depends upon the relative mass fractions of these two components and the importance of the radiative cooling. It is generally observed to be of order $kT_e \sim 1$ keV.

At this temperature, the sound-crossing timescale for the hot gas is only a few million years, while the cooling time is much longer, of order 10^8 years for the density of 0.01cm^{-3} implied by the X-ray properties of a typical galaxy. In this case, if the hot gas is bound to the galaxy, it can be assumed to be near its hydrostatic equilibrium:

$$\nabla P_H = -\rho_H \nabla \Phi(r) \tag{7.55}$$

where the hot gas has a pressure and density of P_H and ρ_H, respectively, and $\Phi(r)$ is the gravitational potential of the galaxy. For spherical symmetry, this becomes:

$$\frac{1}{\rho_H} \frac{dP_H}{dr} = -G \frac{M(r)}{r^2} \tag{7.56}$$

This has an important observational consequence. What we can observe is the radial distribution of the surface brightness and the X-ray spectrum of the hot gas as a function of radius. If we fit these to models of the CIE emissivity working in from the outside, then we can then reconstruct $\rho_H(r), T_e(r)$ and hence $P_H(r)$. These can then be used in the hydrostatic equation (7.56) to obtain $M(r)$, which includes the contribution to the mass of the visible stars as well as the contribution of whatever dark matter components happen to be present. In this way X-ray observations provide a powerful technique to determine the mass inventory of the Universe.

These models also provide the means of calculating the mass of the hot component. The typical values found are in the range $8 < \log M_\odot < 11$, which is, in most cases comfortably lower than the mass of the galaxy, and the self-potential of the hot gas can be ignored in equation (7.56).

A useful approximation to the mass distribution for early-type galaxies is the analytical approximation to an isothermal distribution by King, for which the local (star + dark) matter density is given in terms of the core radius, r_{core}, by:

$$\rho(r) = \rho_0 \left(1 + x^2\right)^{-3/2} : x = r/r_{core} \tag{7.57}$$

which provides a potential distribution:

$$\Phi(r) = -4\pi G \rho_0 r_{core}^2 \frac{\ln \left[x + \left(1 + x^2\right)^{1/2}\right]}{x} \tag{7.58}$$

The central density ρ_0, the core radius and the velocity dispersion are all related:

$$\sigma^2 = (4\pi/9)\, G\rho_0 r_{core}^2 \qquad\qquad (7.59)$$

so that the central potential is $\Phi_0 = -9\sigma^2$. This implies that if the temperature of the hot gas exceeds about 1.8×10^7 K ($kT_e \sim 1.4$ keV) then the gas would no longer be bound, and then would flow out in a thermal wind into the cluster or intercluster medium. In fact, the models of Ciotti *et al.* (1991) show that this can happen in the early phases when the energy input exceeds the ability of the gas to cool, but this evolves to a subsonic flow and finally develops into a cooling flow as the ratio $\dot{E}_{mech}$ / $\dot{E}_{cool}$ decreases during the ageing of the stellar population. Paradoxically, models with a strong wind are much less luminous in X-rays than models with a cooling flow, since the wind models are non-radiative, and the energy put in by supernovae goes into mechanical energy in the outflow.

7.6.2 Clusters of Galaxies

Virtually all of the discussion in the previous section carries over to the case of hot gas in clusters of galaxies. However, the typical velocity dispersions of the clusters are about 1000 km s^{-1}, so the potential is much deeper, and gas temperatures of 5×10^7 K ($kT_e \sim 4$ keV) are typical. In this case we are clearly into the electron Brehmstrahlung-dominated zone of the cooling function, and the cooling timescale is $\tau_{ff} \sim 5 \times 10^{10} n_{-3}^{-1} T_8^{1/2}$ yr, with the density and temperature in units of 10^{-3} cm^{-3} and 10^8 K, respectively. In most cases, the cooling timescale is generally longer than the age of the Universe.

Using the *Einstein* X-ray observatory (Sarazin, 1988) the cluster luminosity is found to lie in the range $39 < \log L_X < 43$ while the mass of X-ray gas is estimated to cover the range $12 < \log M_\odot < 14.5$, (assuming $H_0 = 75$km s^{-1} Mpc^{-1}). These parameters nicely extend the range observed for single galaxies.

The hot cluster gas is metal-rich, and so must have been derived from the stellar populations of the cluster members. This can occur in four ways. First, the gas may have come from the hot galactic winds of individual cluster members. Second, it may be derived from stars which have been tidally stripped from galaxies in the cluster. It is known, for example, that an important intra-cluster population of PNe exists in the Virgo cluster. Thirdly, some portion may have been stripped from individual galaxies by the ram-pressure of the intra-cluster hot gas. In this case, if the orbital velocity of the galaxy in the cluster is v_{gal}, and the pressure in the hot medium in the vicinity of the galaxy is P, the galaxy will be stripped of its hot halo down to a radius r, given roughly by $P \sim \rho_{gal}(r)v_{gal}^2$. Clearly, galaxies that fall in plunging orbits towards the cluster centre will be more likely to be stripped by this process. Lastly, it may have been ejected from individual galaxies by the result of the interaction of relativistic jets of gas from an active galactic
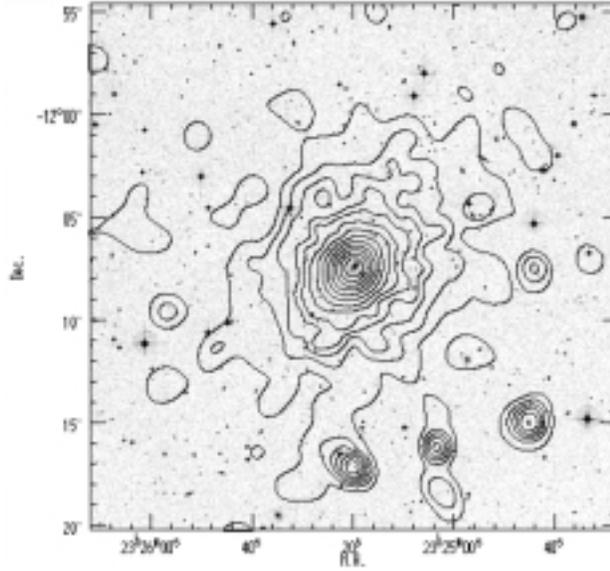
Fig. 7.5. Contour plot of the X-ray surface brightness of the central 4.4 Mpc of the A2597 cluster, superimposed on the optical image. This is a cooling-flow cluster, with a mass inflow of about 350 solar masses. There are more than 10^{14} solar masses of hot gas. The data were obtained with the ROSAT X-ray satellite by Sarazin & McNamara (1997).

nucleus (AGN) with the interstellar medium of the galaxy. We return to this point later in Section(8.3.3).

Because the density of the hot gas is directly related to the depth of the potential, it may be that the cooling timescale in the cluster core becomes shorter than the dynamical timescale, even though in the outer regions of the cluster the cooling timescale is much longer than a Hubble time. In this case, the pressure term in the hydrostatic equation (7.56) is removed, and a stationary solution is no longer possible. Instead a *cooling flow* develops in the central region of the cluster, which is not only bright in X-rays, but may also become prominent at optical wavelengths as the temperatures fall to a point where recombination of hydrogen can occur.

Provided that the change in the gravitational potential is unimportant compared with the energy lost in the cooling, then the mass flux in the cooling flow, $\dot{M}$, can be estimated from the X-ray luminosity of gas at any particular temperature, since this simply reflects the loss of the thermal energy of the gas:

$$\Delta L\left(T_e\right) = \frac{5k\Delta T_e\ \dot{M}}{2\mu m_H} \tag{7.60}$$

However, for the optical flux, such estimates break down. Since each hydrogen atom involved in a cooling flow must ultimately recombine, the recombination flux should, in principle, provide an estimate of the mass flux in the cooling flow ($M_\odot$ yr^{-1}). In practice this is difficult because each hydrogen atom involved in the flow makes a number of recombinations, and collisional excitations also increase the Balmer line flux. The multiple recombinations are caused by two effects. First, the cooling instabilities generate secondary (re-heating) shocks which re-ionise hydrogen atoms possibly more than once, and second, the EUV photons generated in the cooling plasma (particularly below 10^7 K), are produced very close to the recombination zone of the flow, and so are efficiently trapped by the cooler gas. This produces many photoionisations, each of which leads to another recombination. The Balmer flux may thus be enhanced by a factor up to 10 to 30 times over what would be expected on the basis of a single recombination. With this correction factor, the mass fluxes in the cooling flows agree with those derived from the X-ray data, and may be considerable (up to about 500 $M_\odot$ yr^{-1}!). The problem then arises, what happens to all of this gas which is dropping out of the hot phase in and around the central galaxy; which is usually of the massive cD type. It has been argued that this may be forming many new low-mass stars in the central galaxy. Some of it may also find its way down the core of the central galaxy and be used to feed the massive black hole which lurks there.

If the local input of energy is greater than the radiative losses, and the energy of injection is greater than the binding energy, then a *cluster wind* will develop instead. This is analogous to the thermally-driven *Parker winds* which may occur from stars. Provided that the flows are subsonic and spherically symmetric, both winds and cooling flows can be described by the equations of continuity, momentum and the energy conservation. These equations must be solved numerically, and a wide variety of models have been generated in this way.

Apart from the production of X-rays, the existence of a hot cluster medium can be inferred in at least three other ways; through observations of ram-pressure stripping of cluster galaxies, through the influence of the cluster medium on the cosmic microwave background, and through the frequency dependence of the Faraday rotation it induces in linearly polarised radio sources seen through the cluster.

Gas-rich disk galaxies (spiral or irregulars) moving in the cluster medium will experience ram-pressure stripping down to a radius r wherever the ram pressure of the hot gas on the galaxy exceeds the gravitational force binding the gas to the host galaxy:

$$\rho_H v_\perp^2 > 2\pi G \sigma_T\left(r\right)\sigma_g\left(r\right) \tag{7.61}$$

where $\sigma_T\left(r\right)$ is the total surface density in the galactic disk at radius r, $\sigma_g\left(r\right)$ is the gas surface density, ρ_H is the density in the hot medium, and $v_\perp$ is the component of the galaxy's motion through the cluster along the

perpendicular to the disk. Graphical observational evidencence that such ram pressure stripping occurs has been obtained by Cayatte *et al.* (1989), who find that in the Virgo cluster, the H I disk sizes of cluster spiral galaxies becomes systematically smaller the closer they are to the cluster centre. In addition, Giovanelli & Haynes (1985) observed the H I content of galaxies in a number of clusters, and found that the fraction of galaxies displaying an H I deficiency for their Hubble type increased as the X-ray luminosity of the cluster increases.

Even though the cluster gas is of very low density, radiative transfer effects through it may become important. For example, a massive cooling flow has a Thompson electron scattering optical depth given by equation (7.3)). This is of the order 10^{-2} . The hot gas is therefore capable of scattering as much as a few percent of the visible starlight. In addition, this finite optical depth is important in scattering the *cosmic microwave background* (CMB) photons which are the relict of the Big Bang. Because the electrons in the cluster are much hotter ($\sim 10^8$ K) than the radiation temperature of the CMB (2.7 K), they heat the CMB photons, leading to a diminuntion of the CMB at radio wavelengths. This is called the *Sunyaev-Zel′dovich effect*, and it provides a very powerful set of cosmological tests.

Finally, the magnetic field in the cluster produces an anisotropy in the speed of transmission of radio radiation through it. That is, the electron medium with a directional magnetic field is birefringent and is capable of rotating the plane of polarisation of linearly polarised radio sources embedded within or behind it, such as the synchrotron emission from Active Galactic Nuclei or the jets which they produce. The angle of this *Faraday rotation*, ϕ, is:

$$
\begin{aligned}
\phi \;\; &= R_m \lambda^2 \\
R_m &= \frac{e^3}{2\pi m_e^2 c^4} \int n_e B_\parallel dl
\end{aligned}
\tag{7.62}
$$

where $B_\parallel$ is the component of the magnetic field parallel to the direction of propagation of the radiation. If the path length through the medium, l, is measured in parsecs, the density in cm^{-3} and the field in μG, then the rotation measure R_m is 0.815 rad.m^{-2}, so a cluster may produce a rotation measure of a few hundred. The very bright and extended double-lobe radio source Fornax A provides an ideal source with which to study the spatial variations of the product of magnetic field and density provided by Faraday rotation measurements.

7.6.3 Disk Galaxies

Although some disk galaxies also appear to contain both hot and cooling gas, the physics of the heating of this material, and its escape from the disk to form a hot corona is physically rather different from the cases described

above. This is because the energy input in disk galaxies is derived mostly from the winds of hot, young stars and their supernova explosions, and the heated gas interacts strongly with the dense interstellar medium of the galaxy before escaping into the halo.In addition, the systematic rotation of the disk can be an important factor.

Shapiro & Field (1976) first suggested that the material which escapes into the halo does so by passing through buoyant and over-pressurised *bubbles, chimneys*, or *"worm-holes"*. As a result, it flows up and away from the galactic plane until such time as it cools, loses buoyancy and falls back to the disk in dense blobs and filaments to complete the *"Galactic fountain"*.

In our own Galaxy, there is a lot of observational evidence revealing the existence of a highly-ionised component. This has been recently reviewed by Savage (1995). From H I observations, or from UV absorption measurements along the lines of sight to stars at high galactic latitude, the density profile of the un-ionised gas in the solar neighbourhood can be established (Bohlin, Savage & Drake 1978; Diplas & Savage 1994; Dickey & Lockman 1990). Combining the results of these works, the gas density profile can be fitted with two components; the so-called *cold neutral medium* (CNM) and the *warm neutral medium* (WNM):

$$\left\langle n_{\mathrm{H\,I}} \right\rangle \sim 0.4 \exp\left[-\frac{z^2}{2h_{CNM}^2} \right] + 0.16 \exp\left[-\frac{z}{h_{WNM}} \right] \ \mathrm{cm}^{-3} \qquad (7.63)$$

where the (gaussian) scale height of the CNM, h_{CNM}^2 is about 110 pc, and the (exponential) scale height of the WNM is about 360 pc.

The presence of a *warm ionised medium* (WIM), sometimes called *diffuse ionised gas* (DIG), or, more confusingly, the *Lockman-Reynolds layer* can be inferred either directly from observations of diffuse Hα emission, or through observations of pulsar dispersion measures. These two techniques give a different weighting to the material along the line of sight, since the Hα emission measure depends on $\left\langle n_e^2 l \right\rangle$ while the time delay between two frequencies induced in the pulsar signal, $\Delta\tau_D$, depends on $\left\langle nl \right\rangle$;

$$\Delta\tau_D = \frac{e^2}{2\pi m_e c} \left[\nu_1^{-2} - \nu_2^{-2} \right] \int_0^l n_e dl \qquad (7.64)$$

where the integral is called the *dispersion measure* (DM) (cm^{-2}). The pulsar dispersion measurements should give the most reliable result since the quantity we are most interested in is being directly measured, and since the distances of the pulsars can be accurately determined. The mean z - distribution of matter is determined from the component of dispersion measure perpendicular to the galactic plane, DM.sin$|b|$. Reynolds (1993) finds that a two-component fit is best, while Nordgren, Cordes and Terzian fit a simple exponential:

$$\langle n_{\mathrm{H\,II}} \rangle \sim 0.022 \exp\left[-\frac{z}{h_2}\right] \; \mathrm{cm}^{-3} \qquad (7.65)$$

with $h_2 = 670$ pc. This ionised gas is probably at about 10^4K and is kept ionised by the combination of H II regions and supernovae in the disk, although stars cooling to become White Dwarfs should be more important sources of ionising radiation at high scale heights. While these relations establish the distribution of the major gaseous components of the disk, they do not say anything about the hot gas which may be present. Indeed, they are not even able to establish the pressure of this component, even if is in approximate pressure balance with the WNM+WIM, since the local density of these components may be much higher than suggested by (7.63) and (7.65), depending on the (unknown) *volume filling factor* of these components. Reynolds (1980) estimated $P/k \sim 3000$ cm^{-3} K from a study of high-latitude Hα emission, however the total pressure may be a good deal higher than this once the pressure of the magnetic field is taken into account.

The existence of a truly hot component can be directly established through observations of a diffuse X-ray component, or through UV absorption line observations of highly ionised species in the line of sight to halo stars. The diffuse X-ray measurements are confusing because it is uncertain how much of the observed flux comes from the local hot bubble of gas associated with local massive star formation, or how much comes from unresolved X-ray sources at cosmological distances (AGN and the like). However, UV absorption measurements reveal interstellar absorption of Si IV, C IV, N V and O III which can only have come from a hot component. In general, the higher excitation species seem to be confined closer to the plane of the disk. In the column of gas in front of the QSO 3C273, the only extragalactic object where O VI has so far been detected (using the *Hopkins Ultraviolet Telescope*); $\log\left[N(\mathrm{O\,VI})/N(\mathrm{C\,IV})\right] \sim -0.1$. For the stars in the halo, $\log\left[N(\mathrm{O\,VI})/N(\mathrm{C\,IV})\right] \sim 0$, while for stars in the disk, $\log\left[N(\mathrm{O\,VI})/N(\mathrm{C\,IV})\right] \sim 0.8$ (Spitzer, 1996). In addition, Sembach & Savage (1992) find that the scale height of the gas derived from N V absorption measuements is about 1.6 kpc, while in C IV absorption the scale height $\sim$ 5kpc.

Even though these species are highly ionised, they cannot arise in gas as hot as 10^6 K. With some dependence on their mode of ionisation, most likely come from a gas of 1 - 3) $\times 10^5$ K. This presents something of a problem to models, since this is very near the peak of the cooling curve, where the cooling timescales are very short. The highly ionised species must therefore arise in a layer which has a continuous source of energy input, and with a rather generous energy budget. Since about one supernova explosion occurs in our Galaxy roughly once every 30 years, and each liberates about 10^{51} ergs of kinetic energy, then the energy required by the hot halo gas is up to about 10% of the energy produced by all the supernova explosions in the galaxy.

The mechanism of this energy input is still rather uncertain. Shull and Slavin (1994) advocate turbulent mixing layers and isobarically-cooling supernova remnants to provide a mass flux of about 20-40 $M_\odot$ yr^{-1} and $\sim 10^{41}$ erg s^{-1} into the halo. Shapiro and Benjamin (1993) suggest that cooling gas is by itself able to explain the absorption line observations, provided that the absorption of the ionising photons produced in the cooling flow is properly taken into account, while Spitzer (1996) proposes that the highly ionised species arise in conduction fronts and turbulent mixing layers.

The energy balance of the Galactic corona was discussed in a classical paper by Chevalier & Oegerle (1979). If the gas is in hydrostatic equilibrium, then we must solve the hydrostatic equation in cylindrical coordinates, taking account of the centrifugal term due to the rotation of the disk, *c.f.* equation (7.56);

$$\frac{1}{\rho_H} \frac{dP_H}{dz} = -\frac{d\Phi}{dz}$$

$$\frac{1}{\rho_H} \frac{dP_H}{dr} = -\frac{d\Phi}{dr} + \frac{v_{rot}^2}{r} \tag{7.66}$$

where v_{rot} is the local rotational velocity. Presumably, the gas is initially rotationally supported, since it is produced by the rotationally supported stellar population of the disk. However, it has an additional thermal energy component, and as it rises into the hot halo, it will also expand radially. The hot gas in the halo should therefore rotate more slowly than the disk.

In order to escape from the Galaxy as a *galactic wind*, the gas has to overcome the gravitational potential of the galaxy, so that, if it is ejected from the plane an initial radius r_0 at an intial velocity of v_0, it must have an initial temperature higher than the critical value, T_{crit} :

$$T_{crit} > \frac{\gamma - 1}{\gamma} \frac{m_H}{k} \left(\Phi\left(r_0, 0\right) - \frac{v_{rot}^2}{2} - \frac{v_0^2}{2} \right) \tag{7.67}$$

If the inital velocity of ejection is small compared with the rotation velocity, and the gravitating mass is centrally condensed then the term in brackets is of order $v_{esc}^2(r_0)/2$. This would imply a critical temperature of about 10^6 K in the solar neighbourhood, and about 4×10^6 K in the Galactic centre, above which, in the absence of significant cooling, a wind will be driven.

For our Galaxy, the parameters estimated for the galactic halo ($P/k \sim$ 3000 cm^{-3} K, $\dot{M} \sim$ 20-40 $M_\odot$ yr^{-1} and $\dot{E} \sim 10^{41}$ erg s^{-1} appear to preclude the possibility of a galactic wind. However, there is now a deal of evidence that starburst galaxies galaxies do produce winds, either because their low mass in the case of dwarf galaxies, or because a high energy production through supernova explosions allows them to satisfy equation (7.67). For the dwarf galaxies, an important fraction of the nucleosynthesis products from supernova explosions may be carried away into intergalactic space by this

process, and so dwarf galaxies may be significant contributors to the intra-cluster medium at the current epoch.

For our Galaxy there is evidence that a tenuous hot phase is found at very large distances from the galactic plane. This is furnished by the Magellanic Stream (Mathewson & Ford, 1985), which is an enormous trail of H I clouds streaming out along the orbital path traced out by the Magellanic Clouds. The most compelling theory of its origin is that it is gas removed by ram pressure stripping of the hot halo as the two Magellanic clouds plunged in towards the Galaxy in their orbit over the last few hundred million years (Meurer *et al.* 1985; Moore and Davis, 1994). Currently, the Magellanic Clouds are almost at perigalacticon, and are moving at about 300 km.s^{-1}, so that the ram pressure is at its greatest. Faint X-ray emission has been detected with the ROSAT satellite (Wang & McCray) and Hα emission has been detected at the cloud leading edges, which allows an estimate of the halo density. This is of order 10^{-4} cm^{-3} at a distance of 50 kpc from the Galactic centre (Weiner & Williams, 1996).

Notes on Chapter 7

- Equilibrium and non-equilibrium cooling functions were given by Sutherland, R.S. & Dopita, M.A. 1993, ApJS, 88, 253.
- The best source of information on thermal conduction remains Spitzer, L. Jr., 1956, *The Physics of fully Ionised Gases, 2nd. ed.* J.Wiley Interscience: NY.
- The theory of turbulent mixing layers was developed by Slavin, J.D., Shull, J.M. & Begelman, M.C. 1993, ApJ, 407, 83.
- An up to date account of the theory of thermal instabilities is to be found in Balbus, S.A. 1995, in *The Physics of the Interstellar Medium and Intergalactic Medium*, ASP Conf. Ser. v80; eds A. Ferrara, C.F. McKee, C.Heiles & P.R. Shapiro, ASP:SanFrancisco, p328 (ISBN 0-9377-7-99-6). This volume also contains very useful reviews on the galactic corona (Savage, B.D. p233) and the hot ISM in galaxies (Fabbiano, G., p419), all of which have been used extensively in the preparation of this chapter.
- For further reading on the subject of hot gas in clusters of galaxies see Sarazin, C.L. 1988, *X-ray Emissions from Clusters of Galaxies,* CUP:Cambridge (ISBN 0-512-32957-4). For the observational material on the X-ray haloes of galaxies, see Fabbiano, G., Kim, D-W., & Trinchieri, G. 1992, ApJS, 80, 531.
- A classical and most readable paper on the physics of the hot gas in the Galactic halo is by Chevalier, R.A. & Oegerle, W.R. 1979, ApJ, 227, 398. The observational material is best summarised by Savage, B.D. in *The Physics of the Interstellar Medium and Intergalactic Medium*, ASP Conf. Ser. v80; eds A. Ferrara, C.F. McKee, C.Heiles & P.R. Shapiro, (ASP:SanFrancisco), p233 (ISBN 0-9377-7-99-6).

Exercise 7.6.1. Assume that an astrophysical plasma consists (by number of atoms) of 90% hydrogen, 9% helium and just 1% heavier elements with a mean atomic number of 8.

a. Calculate the mean atomic weight of the plasma in the limits when it is un-ionised, and when it is fully-ionised.

b. What is the mean thermal energy per particle (ions and electrons both) of this plasma at a temperature T, assuming that the plasma is fully ionized?

c. What is the total thermal heat content of the plasma per unit volume at a temperature T and a particle density n?

d. Assuming that the plasma cools by the electron bremssrahlung (free-free) process with a cooling rate given by $\Lambda = 3 \times 10^{-27} T^{-1/2}$ ergcm^{3}s^{-1}, provide an exact expression for the characteristic cooling timescale of the plasma at the temperature T and particle density n.

e. Provide an exact expression for the time taken for this plasma to cool at constant pressure down to absolute zero, assuming that the gas can remain fully ionized and cools only by the bremssrahlung process.

Exercise 7.6.2. In the halo of our Galaxy, there is a warm (~10,000K) ionised medium extending to more than 1kpc above the galactic plane. This is often called the *Reynolds Layer*, named after the researcher who has studied it in most detail. Assume that this medium is heated by photons from starlight, which are absorbed by the gas, and deliver a heating rate to the gas of $\Gamma_g = 10^{-24} n^2$ erg cm^{-3}s^{-1}, and which are also absorbed by the dust, and deliver a heating rate of $\Gamma_d = 10^{-26} n$ erg cm^{-3}s^{-1} to the gas. The cooling rate of the gas can be approximated by $\Lambda = 10^{-24} n^2 T_4$ erg cm^{-3}s^{-1}, where T_4 is the temperature of the gas measured in units of 10,000K. Assuming that the Reynolds Layer has a density distribution $n(z) = 0.1 \exp[-z/1\text{kpc}]$ cm-3, where z iis the height above the Galactic midplane, calculate an expression for the temperature distribution as a function of height above the plane.

Exercise 7.6.3. A (highly idealized theoretician's) spherical galaxy has a total mass M, a total radius R_0, and a stellar density distribution which varies as r^{-2}. These stars are in circular orbits with random angular momentum about the centre, giving no net rotation to the system. Gas is blown off stars in the galaxy with a very low initial velocity relative to the stars, and this gas is heated by cloud-cloud collisions to coronal temperatures.

a. What is the velocity dispersion of the stars as a function of radius?

b. Assuming $M = 5.0\text{x}10^{11} M_\odot$ and $R_0 = 10$ kpc (1pc $= 3.08\text{x}10^{13}$ km; solar mass $= 1.99\text{x}10^{30}$ kg), what is the average relative velocity of the stars as a function of radius? If the hot gas contains 10% by number of helium atoms, negligible numbers of heavier atoms and is fully ionized, to what temperature is the gas heated, again as a function of radius?

c. If the hot gas is in hydrostatic equilibrium and fully ionized at all radii, and the density at R_0 is 10^{-2} cm^{-3}, what will be the hydrogen density at a radius of 1 kpc?

d. Assuming that the gas cools at a rate $= 10^{-22}(T/10^6\mathrm{K})^{-1/2}n_H^2$ erg cm^{-3}s^{-1}; what is the cooling timescale of the hot gas at 10 kpc and at 1 kpc? Comment on the validity of the assumptions made in part (1.3), above.

Exercise 7.6.4. 1 The cooling function of a plasma at low temperature is dominated by IR lines which, taken together, can be taken to be only weakly dependent on temperature. Taking $\Lambda = 10^{-24}$ erg.cm^{3}s^{-1}, estimate the radius below which a cool cloud will be destroyed by thermal conduction in the saturated regime.

8. Interstellar Shocks

"Awaiting the sensation of a short sharp shock
From a cheap and chippy chopper on a big black block"
— W.S. Gilbert (The Mikado, 1885)

Violent events occur everywhere in the interstellar and intergalactic medium. Stars explode in supernova events, newly born stars blow energetic winds, or produce UV photons to ionise and heat their surroundings creating a region of overpressure in the ISM which produces a violent expansion into the lower pressure gas beyond. In the far reaches of space, galaxies and even whole clusters collide with one another, and the active nuclei of individual galaxies spew out gas at velocities approaching the speed of light. All of these events create shock waves, which are the major source of the hot plasma discussed in the previous chapter. In this chapter we will attempt to come to grips with the basic physics of such shockwaves.

8.1 Why Do Shocks Exist?

Fluid dynamical motion is governed by the conservation of mass, momentum, and energy. In addition, the Maxwell equations must also apply, and the flow is also subject to an equation of state which relates the pressure to the other thermodynamic variables; density and temperature. Let us first restrict our attention to flows without magnetic fields, and to flows which are also radiationless and thermodynamically reversible, so that the entropy is constant throughout the flow. All such flows must satisfy the conservation of mass and of momentum. The *continuity equation* which describes the conservation of mass is written in its general vectorial form as:

$$\frac{\partial \rho}{\partial t} + \nabla.(\rho \mathbf{v}) = 0. \tag{8.1}$$

This simply states that the rate of change in the local density is determined by the difference of the rate of flow of material into a given test volume, and out of the same test volume, along the direction of motion. For a flow which is restricted to one dimension, this simplifies to:

$$\frac{d\rho}{dt} + \frac{d}{dx}(\rho v) = 0. \tag{8.2}$$

Further simplification occurs for *steady flows, i.e.* flows which are invariant with respect to time. For these:

$$\frac{d}{dx}(\rho v) = 0. \tag{8.3}$$

Turning now to the next conservation equation, that of momentum. In vectorial form this is:

$$\rho \left(\frac{\partial \mathbf{v}}{\partial t} + (\mathbf{v}.\nabla)\mathbf{v} \right) = \mathbf{F} - \nabla P. \tag{8.4}$$

This equation, the *Eulers's Force Equation* states that the rate of change of momentum of material within a test volume is equal to the flux of momentum into that volume plus the forces acting upon this material. These are composed of both external forces such as gravitational forces, *etc.*, and internal forces resulting from the pressure gradient. For a one dimensional flow which is not subject to any external forces, this simplifies to:

$$\rho \frac{dv}{dt} + \rho v \frac{dv}{dx} = -\frac{dP}{dx}. \tag{8.5}$$

This equation simplifies still further for steady flows. Again, the time derivative vanishes, and since, from the continuity equation ρv is a constant, we now obtain:

$$\frac{d}{dx}(P + \rho v^2) = 0. \tag{8.6}$$

Finally, to complete the description of a one dimensional, thermodynamically reversible, steady flow, we require the (adiabatic) equation of state,

$$P = K\rho^\gamma \tag{8.7}$$

where K (a function of the entropy) is a constant, and where $\gamma = C_P/C_V$ is the ratio of the specific heats at constant pressure and at constant volume. In a monatomic gas, $\gamma = 5/3$, is 7/5 for a diatomic molecular gas such as H_2, and is 4/3 for a polyatomic gas with many internal degrees of freedom.

Now, consider what happens when a small disturbance is propagated in an (initially steady) one-dimensional flow. Assume that the fluid is initially at rest, that the change in velocity is small and that the change in the other flow variables ρ, P is small compared with their initial values ρ_0, P_0. In this case we can neglect the second term on the left hand side of equation(8.5), since it is the product of two small quantities. Likewise we can neglect the second part of the expansion of $d(\rho v)/dx$ in the continuity equation (8.2), so that:

$$\frac{dv}{dx} = -\frac{1}{\rho_0}\frac{d\rho}{dt}. \tag{8.8}$$

Furthermore, from the equation of state (8.7) we can write the pressure change in terms of the density change:

$$dP = \left(\frac{\gamma P_0}{\rho_0}\right) d\rho. \qquad (8.9)$$

This gives, when subsituted into the *linearised* momentum equation (*i.e.* the momentum equation obtained by retaining only the first order quantities),

$$\frac{dv}{dt} = -\frac{1}{\rho_0}\left(\frac{\gamma P_0}{\rho_0}\right)\frac{d\rho}{dx}. \qquad (8.10)$$

Differentiating equation (8.8) with respect to time, and equation (8.10) with respect to distance, and subtracting the two resulting equations, we have finally;

$$\frac{d^2\rho}{dt^2} + \left(\frac{\gamma P_0}{\rho_0}\right)\frac{d^2\rho}{dx^2} = 0 \qquad (8.11)$$

This should be recognised as the wave equation. It implies that changes in density are propagated in the fluid at a certain velocity (the speed of sound), c_s:

$$c_s = \left(\frac{\gamma P}{\rho}\right)^{1/2} \qquad (8.12)$$

Thus, if a change is made at one place and time, such as a sudden local increase or decrease in pressure, the effects can propagate relative to a given fluid element at the sound speed.

As a consequence, analogous to probelems involving causality in special relativity, we can define a time-like region, within which points are connected causally through the propagation of sound waves. In this region, the local fluid variables can change in response to the disturbance. There is also a space-like region within which no change can occur; *see* figure (8.1). In x, y, t -space, the time-like region is a cone and in x, y, z, t -space the time-like region is an (expanding) sphere. As figure (8.1) shows, motion of the medium tips the cone, so that if the velocity of the medium exceeds the speed of sound, information about the change can never be carried backward in the spatial (*Eulerian*) x -coordinate.

Now, suppose the change which is being propagated is an adiabatic compression of the medium which increases the density from an initial density ρ_1 to ρ_2. Such an adiabatic compression obeys $P = K\rho^\gamma$ so that equation (8.12) implies that, following the compression, the speed of sound in the medium is increased in the ratio $(\rho_2/\rho_1)^{(\gamma-1)/2}$, or in the ratio $(\rho_2/\rho_1)^{1/3}$ in the case of a monoatomic gas. As a result, the most compressed region has a tendency to catch up with and to overrun the uncompressed region. This clearly has the potential to create a discontinuity in the variables of the fluid flow.

To consider this effect in more detail, suppose that the object creating the compression is a piston which has been smoothly accelerated from rest to
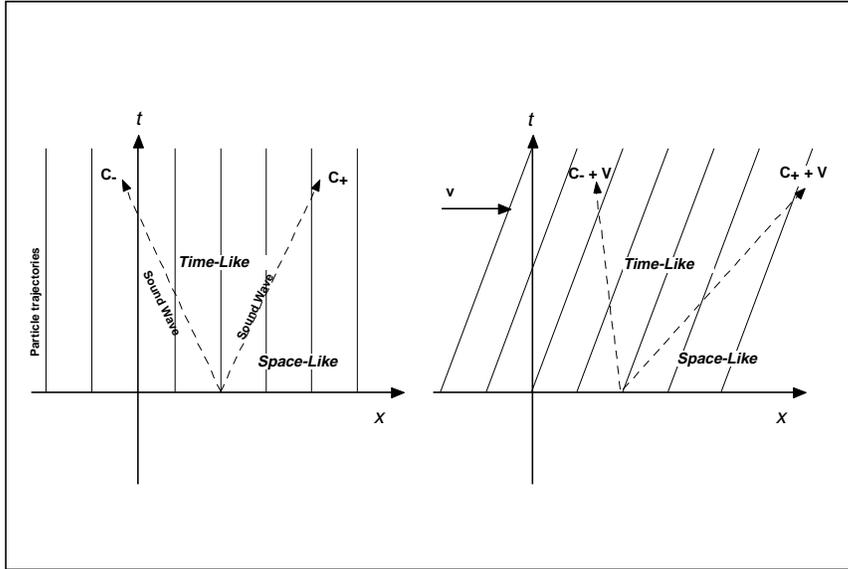
Fig. 8.1. In a fluid, information about changes in conditions is carried by sound waves from place to place. In a stationary medium (left), points in the time-like cone are accessible to one another, while points in the space-like region are not. In a moving medium (right) the time-like cone is tipped over. When the velocity of the medium exceeds the speed of sound, information cannot propagate backwards in the spatial (Eulerian) coordinate.

some high velocity, greater than the sound speed in the undisturbed medium as in figure (8.2). Initially, sound waves can be carried forward at a speed which is faster than the speed at which the piston is moving, and a smooth gradient of density is therefore set up ahead of the piston. However, with time and with the acceleration of the piston, the forward propagating sound waves converge closer and closer until at a certain point they merge together. At this point in the flow, a discontinuity in the flow variables (a shock) develops, and the information which was carried by the separate sound waves is lost. This information (in principle) would have allowed us to reconstruct the details of the acceleration of the piston. Loss of information equates to an increase of the entropy, and therefore the development of a shock is an irreversible process.

If instead, the piston were to be withdrawn, the fluid variables remain continuous throughout the flow. This is shown in figure (8.3). In this case the particle trajectories diverge. Since the work done by the adiabatic expansion of the gas causes it to cool, lowering the sound speed, the forward- and backward- propagating soundwave trajectories tend to approach the particle trajectories asymptotically. If the piston is withdrawn too rapidly, the internal
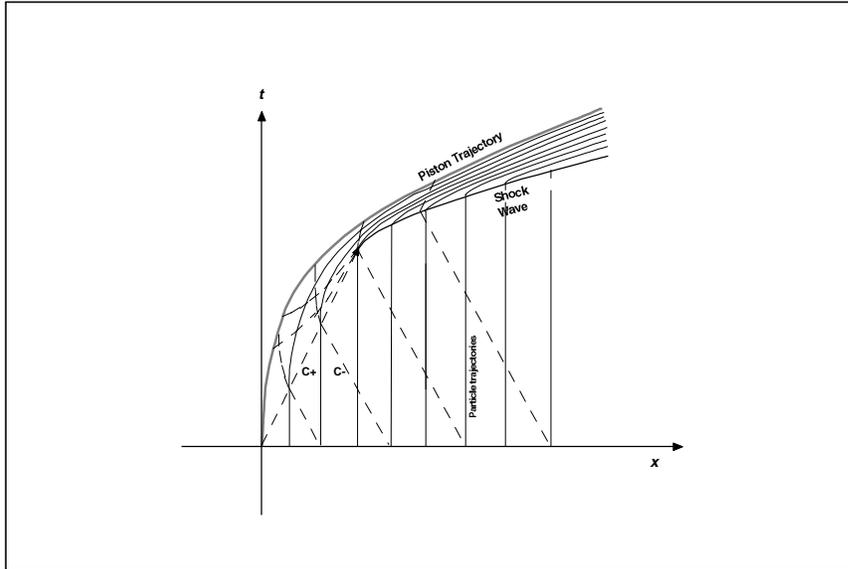
Fig. 8.2. A sketch of the particle trajectories and the paths of sound waves for a smoothly accelerated piston. The increase of sound speed with density leads first to a convergence of sound waves, and then to the development of a shock discontinuity in the flow, in which information is lost and entropy is therefore increased. The shock gradually thickens as more and more gas is swept up.

energy of the gas is depleted entirely. The maximum expansion velocity of the gas is $c_1 \left[2/\left(\gamma - 1 \right) \right]^{1/2}$, where c_1 is the speed of sound in the undisturbed plasma. If the piston is withdrawn faster than this, a vacuum develops behind the piston. The disturbance which results from an expansion such as depicted in figure (8.3) is termed a *rarefaction wave*.

8.2 J-Shocks

The sketch presented in Figure (8.2) suggests that in the shock which develops in a compression wave, the flow variables change in a discontinuous way. In practice, this is effectively true when the plasma is highly ionised, since in this case the ions and electrons are strongly coupled by the Coulomb interactions, and the effective mean free path of any charged species is very short. Such shocks are referred to as *Jump-shocks (J-Shocks)* to distinguish them from shocks in which some of the flow variable change in a more continuous manner over space, which are called *Continuous or C-Shocks*.

Across the shock, the equation of continuity and momentum conservation apply. The momentum equation is very easily modified to take into account
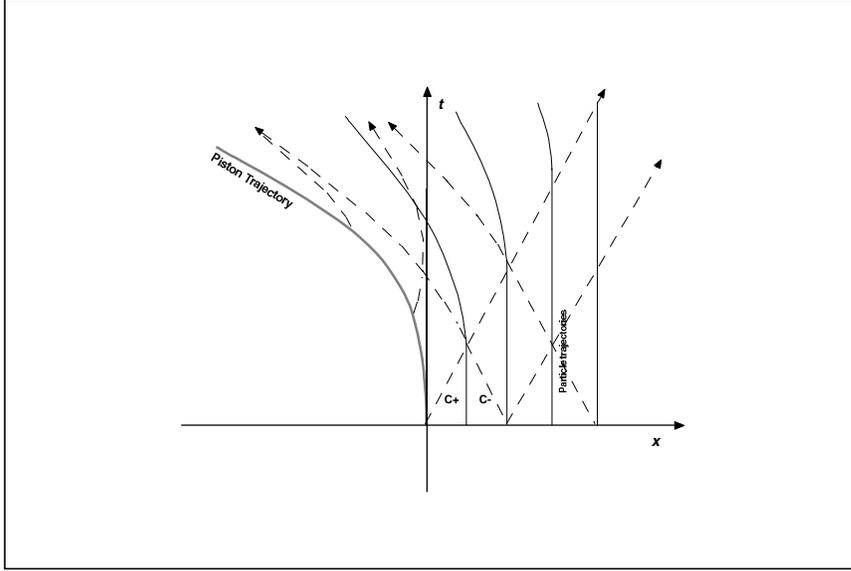
Fig. 8.3. A schematic of the development of a rarefaction wave caused by the withdrawal of a piston. Particle trajectories and the trajectories of soundwaves all diverge, so in this case no discontinuities in the flow can develop.

an internal magnetic field. If the field lies along the direction of motion, it plays no part in the hydrodynamic flow, since it unchanged by the flow and also provides no pressure support to the flow. If, on the other hand, it lies transverse to the flow, we must include the effects of the magnetic field pressure, since the field can be either compressed or rarefied by the flow. If the component of the magnetic field transverse to the direction of flow is B, the momentum equation becomes:

$$\frac{d}{dx}(P + \rho v^2 + \frac{B^2}{8\pi}) = 0. \quad (8.13)$$

If the plasma has a reasonable degree of ionisation so that its conductivity is high, then it is reasonable to regard the magnetic field as being frozen into the flow. In this case we can write an equation of magnetic flux conservation:

$$\frac{d}{dx}(Bv) = 0. \quad (8.14)$$

Across the shock, energy is also conserved. However, in the shock flow as a whole, energy may be lost or gained by the emission or absorption of radiation. In its general form, the law of *Conservation of Energy* requires that the rate of change of energy of the gas within any volume must equate to the net flux of energy through the surface which bounds that volume. Let us write this in generalised vectorial form:

$$\frac{\partial}{\partial t}\left[\frac{\rho v^2}{2} + \rho U + \rho \Phi\right] + \nabla.\left[\rho \mathbf{v}\left(\frac{v^2}{2} + H + \Phi\right)\right] + \nabla.\mathbf{F} = 0 \qquad (8.15)$$

Here the terms in the time derivative represent the kinetic energy, the internal energy, and the potential energy (usually gravitational). $H = U + P/\rho$ is the *Enthalpy* per unit mass (the Specific Enthalpy). This represents the total amount of energy per unit mass available to the gas. The last term represents the divergence of the radiative flux:

$$\nabla.\mathbf{F} = -4\pi\kappa J + n^2 \Lambda \qquad (8.16)$$

where κ is the net absorption coefficient of the gas, and J is the intensity of the local radiation field, so that the first of the two terms on the right hand side of equation (8.16) represents the energy absorbed inside the test volume from the radiation field. Λ is the cooling function defined in Section (7.1), and n is the total particle density, so that the final term in the equation represents the radiative losses from the test volume. (Here, we write these losses as $n^2\Lambda$ rather than as $n_e n\Lambda$ since in the un-ionised molecular or atomic plasmas which we will also wish to consider, electron-particle collisions are not the main radiative loss mechanism). For a one-dimensional, potential-free, steady flow, the time derivatives vanish. Including a term to account for the magnetic field energy stored in the magnetic field transverse to the direction of flow, equations (8.15) and (8.16) reduce to:

$$\frac{d}{dx}\left(\frac{\rho v^3}{2} + (U + P)v + \frac{B^2 v}{4\pi}\right) = 4\pi\kappa J - n^2 \Lambda \qquad (8.17)$$

8.2.1 The Rankine-Hugoniot Jump Conditions

The energy equation (8.17), together with the equation of continuity (8.3), the momentum equation (8.13), the magnetic flux conservation equation (8.14), and the equation of state (8.7), provide a complete description of all one-dimensional, potential-free steady flows, and can be readily integrated to provide relationships between any two regions of the flow. Consider an "initial" point in the flow with hydrodynamical variables v_0, P_0, ρ_0 and magnetic field B_0 and a later point in the flow where these have changed to v_1, P_1, ρ_1 and B_1. If we eliminate the internal energy term from equation (8.17), using the equation of state; $U = P/(\gamma-1)$ we can then write the *Rankine-Hugoniot Jump conditions* in terms of the difference in the fluid variables evaluated at the two points (0 and 1; respectively):

$$\begin{aligned}
[\rho v]_0^1 &= 0\\[4pt]
[Bv]_0^1 &= 0\\[4pt]
\left[P + \rho v^2 + \frac{B^2}{8\pi}\right]_0^1 &= 0 \qquad (8.18)\\[4pt]
\left[\frac{\rho v^3}{2} + \frac{\gamma}{\gamma-1}Pv + \frac{B^2 v}{4\pi} + F\right]_0^1 &= 0
\end{aligned}$$

Provided that the initial conditions are specified, and the energy loss term is calculable, these can then be used to derive the flow variables and the magnetic field at any other point in the flow. Because we are dealing with steady flows, it should be clearly understood that for these conditions to apply, any flow discontinuities (shocks) must be stationary in the frame of reference in which the jump conditions are evaluated, since otherwise we would have to re-introduce time derivatives.

8.2.2 Radiationless Shocks

For simplicity, let us consider for the moment only flows without magnetic fields. In a radiation-free change, such as will occur if the flow variables change suddenly, as in a shock, $[F]_0^1 = 0$. With these approximations, the Rankine-Hugoniot jump conditions simplify still more to:

$$
\begin{aligned}
\left[\rho v\right]_0^1 &= 0 \\
\left[P + \rho v^2\right]_0^1 &= 0 \\
\left[\frac{v^2}{2} + \frac{\gamma}{\gamma - 1}\frac{P}{\rho}\right]_0^1 &= 0
\end{aligned}
\tag{8.19}
$$

Since $\rho_1 = \rho_1 v_0 / v_1$ from the continuity condition, then the momentum condition yields $P_1 = P_0 + \rho_0 v_0 (v_0 + v_1)$. Substituting both of these into the energy condition and collecting terms:

$$
\left(\frac{\gamma + 1}{\gamma - 1}\right) v_1^2 - \left(\frac{2\gamma}{\gamma - 1}\right)\left(\frac{P_0 + \rho_0 v_0^2}{\rho_0 v_0}\right) v_1 + \left(\frac{2\gamma}{\gamma - 1}\frac{P_0}{\rho_0} + v_0^2\right) = 0 \quad (8.20)
$$

This simple quadratic equation can be further simplified by eliminating P_0 in favour of the sound speed *c.f.* equation (8.12), and dividing throughout by v_0^2 to make the equation dimensionless :

$$
\left(\frac{\gamma + 1}{\gamma - 1}\right)\left(\frac{v_1}{v_0}\right)^2 - \left(\frac{2}{\gamma - 1}\right)\left(\left[\frac{c_0}{v_0}\right]^2 + \gamma\right)\left(\frac{v_1}{v_0}\right) + \left(\frac{2}{\gamma - 1}\left[\frac{c_0}{v_0}\right]^2 + 1\right) = 0
$$
$$
\tag{8.21}
$$

In fluid dynamics the ratio of the flow speed to the sound speed is defined to be the *Mach Number* of the flow, $\mathcal{M}$, so that equation (8.21) is further simplified to:

$$
\left(\frac{\gamma + 1}{\gamma - 1}\right)\left(\frac{v_1}{v_0}\right)^2 - \left(\frac{2}{\gamma - 1}\right)\left(\mathcal{M}^{-2} + \gamma\right)\left(\frac{v_1}{v_0}\right) + \left(\frac{2\mathcal{M}^{-2}}{\gamma - 1} + 1\right) = 0. \tag{8.22}
$$

For a sufficiently fast flow, the $\mathcal{M}^{-2}$ terms in this equation can be neglected, and therefore in a monatomic gas with $\gamma = 5/3$ we have the solutions:

$$
v_1 = v_0 : v_1 = v_0/4
$$

The first of these is the trivial solution; nothing happens in the flow. The second solution is the more interesting *strong shock* solution, and implies that the monatomic gas can only be compressed at most by a factor of four in its passage through the shock. The strong shock limit is essentially the condition that the gas pressure in the pre-shock gas be negligible compared with the ram pressure. Since the shock is stationary in the frame of reference we are using, v_0 can be equated to the shock velocity, $v_0 \equiv v_s$. The full solution for the post-shock flow variables in this case, the strong shock limit, is therefore:

$$
\begin{aligned}
v_1 &= v_s/4 \\
\rho_1 &= 4\rho_0 \\
P_1 &= 3\rho_0 v_s^2/4.
\end{aligned}
\tag{8.23}
$$

The kinetic energy in the flow is decreased by the passage through the shock. Energy is conserved, so that the decrease in flow energy must be matched by an increase in thermal energy of the plasma. The specific internal energy of the post-shock gas is:

$$
\epsilon_{int} = \frac{3P_1}{2\rho_1} = \frac{9}{32}v_s^2
\tag{8.24}
$$

Interestingly enough, this is the same as the (rest frame) kinetic energy of the post-shock gas,

$$
\epsilon_{kin} = \frac{1}{2}\left(3v_s/4\right)^2 = \frac{9}{32}v_s^2
\tag{8.25}
$$

Since the equation of state can be written $P/\rho = kT/\mu m_H$ we can obtain the *post-shock temperature* in the strong shock limit from equation (8.23):

$$
T_1 = \frac{3\mu m_H v_s^2}{16k}
\tag{8.26}
$$

$$
\sim 1.38 \text{x} 10^5 v_{100}^2 : \text{(fully ionised plasma)}
$$
$$
\sim 2.90 \text{x} 10^5 v_{100}^2 : \text{(neutral atomic plasma)}
$$
$$
\sim 3.04 \text{x} 10^6 v_{100}^2 : \text{(atomic oxygen plasma)}
$$

Here, numerical values have been given for three paricular cases normalised to a shock velocity, v_{100}, of 100 km s^{-1}. The first and second cases apply to ionised and atomic plasmas with solar composition, respectively, and the third to an atomic oxygen plasma such as we might find in a young supernova remnant of a massive star.

As a result of the vastly increased temperature of the post-shock plasma, the velocity of the post-shock flow is comfortably sub-sonic. From equations (8.23) and (8.12), we have $\mathcal{M}_1 = 5^{-1/2}$ in a monatomic plasma. Because the flow is quite subsonic, its ram pressure is relatively unimportant, so that isobaric cooling represents a good approximation to the subsequent flow.

The presence of a magnetic field does not influence the structure of the shock front. The magnetic pressure is increased at most by a factor of sixteen

across a strong shock, but the gas pressure can increase by an arbitarily large value, depending on the velocity of the shock.

8.2.3 Isothermal Shocks

As soon the gas is heated by the shock, it begins to radiate and cool, so that the full Rankine-Hugoniot conditions must be used to solve for the subsequent flow parameters. However, a useful limiting case occurs when the gas is fully cooled and has returned to its original temperature. This occurs, for example, when the initial and final temperatures are maintained by cosmic ray heating or by photoionisation heating. Such a fully radiative shock is called an *isothermal shock*. In this case the sound speed in the post-shock gas is the same as in the pre-shock gas, and we can use this condition in the place of the energy conservation condition:

$$\left[\rho v\right]_0^1 = 0$$
$$\left[P + \rho v^2\right]_0^1 = 0$$
$$\left[\frac{P}{\rho}\right]_0^1 = 0$$
(8.27)

The equation of state for an isothermal plasma is $P = const\rho$, so that effectively, $\gamma = 1$ at the two control points in the pre-shock and post-shock plasma. The speed of sound is $c_s^2 = P/\rho$. Solving equation(8.27) in terms of the Mach number of the pre-shock flow provides the quadratic equation:

$$\left(\frac{v_1}{v_s}\right)^2 - \left(\mathcal{M}^{-2} + 1\right)\left(\frac{v_1}{v_s}\right) + \mathcal{M}^{-2} = 0.$$
(8.28)

which has the solutions:

$$v_1 = v_s : v_1 = \mathcal{M}^{-2} v_s$$
(8.29)

Again, as in the radiationless shock case, we have the trivial solution and the shock solution. From the shock solution, and using the equation of continuity, it is clear that the maximum compression in an isothermal shock is equal to the square of the Mach number of the shock. The total energy radiated in the shock is given by the energy equation (8.17), which shows that the total shock luminosity per unit area, in the absence of a magnetic field is:

$$\dot{E} = \frac{\rho_0 v_s^3}{2}\left(1 - \mathcal{M}^{-2}\right)$$
(8.30)

this equation is particularly useful in astrophysical applications to relate the total shock luminosity to the other shock parameters.

When there is a (transverse) magnetic field present in the pre-shock medium, the pressure due to the transverse component of the magnetic field

is likely to come to dominate the gas pressure at some point in the post-shock plasma, since the magnetic pressure increases as the square of the density, but, according to equation (8.23), the gas pressure is limited by the ram pressure of the material entering the shock. The implication of this is that, at a certain point in the cooling plasma, the shock switches from being gas-pressure supported to being supported by the magnetic pressure. If the magnetic field pressure in the post-shock gas dominates the gas pressure and the ram-pressure terms, and if the gas and magnetic pressure in the pre-shock gas are negligible compared with the ram pressure (the fast shock limit) then, from the momentum equation, we have:

$$B_1 = 8\pi\rho_0 v_s^2. \tag{8.31}$$

Furthermore, if the magnetic field is also frozen in the flow then equating the pre-shock ram pressure to the post-shock transverse magnetic pressure, we obtain the maximum compression factor attainable in the shock:

$$\frac{\rho_1}{\rho_0} = 2^{1/2}\mathcal{M}_A \tag{8.32}$$

where the *Alfvén Mach Number*, $\mathcal{M}_A$, is defined as the ratio of the shock velocity to the Alfvén velocity defined in equation (7.44); $v_A^2 = B^2/4\pi\rho$. In fact, the jump conditions can be readily solved for an arbitary ratio of magnetic pressure to gas pressure in pre-shock gas (*see* Draine & McKee, 1993; referenced in the notes) to give the general formula for the compression factor in an isothermal shock:

$$\frac{\rho_1}{\rho_0} = 4\left\{2\mathcal{M}^{-2} + \mathcal{M}_A^{-2} + \left[\left(2\mathcal{M}^{-2} + \mathcal{M}_A^{-2}\right)^2 + 8\mathcal{M}_A^{-2}\right]^{1/2}\right\}^{-1} \tag{8.33}$$

In a fully-ionised interstellar medium, the Alfvén velocity in the pre-shock gas is likely (although not certainly) to be of order of the sound speed, c_{II}. This is because turbulence in the ISM tends to wind up any pre-existing field. Suppose that magnetic pressure initially provided no pressure support, then the turbulence is not damped by the presence of the magnetic field. However, as the magnetic field increases (at a given density), the magnetic pressure and the magnetic energy both increase. Eventually, we reach an equipartition value for the magnetic field, where the turbulence is dissipated through Alfvén waves making the medium "magnetically stiff". For a fully ionised plasma of solar abundances, the condition of equipartion of energy densities in the thermal and magnetic components gives $\langle B\rangle^2/4\pi \sim 2.3 n_H k T_0$ where $\langle B\rangle$ is the strength of the local magnetic field, and n_H is the hydrogen particle density. This is effectively the same as the condition $v_A \sim c_{II}$, which gives, from equations (7.43) and (7.44) $\langle B\rangle^2/4\pi \sim 5 n_H k T_0/3$. If these conditions are assumed to hold then the transverse field is related to the pre-shock density by $B_0^2/4\pi \sim n_0 k T_0$, approximately.

8.3 The Drivers of Interstellar Shocks

As we have seen, an interstellar shock requires a compressive piston. Such pistons arise in nature in three ways; from supernova explosions which deposit a large quantity of energetic plasma at a single point in space and time, from radiatively-driven stellar winds originating in hot stellar atmospheres, and from energetic outflows from newly born low mass stars or from the nuclei of active galaxies (where the outflows are relativistic).

Here we will consider the simplified hydrodynamics of each of these three classes of object. This treatment draws heavily on Dyson & Williams (1997), to which book the interested reader is directed for further information. For observational aspects we strongly recommend the book by Lozinskaya (1992) referred to in the notes to this chapter.

8.3.1 Supernova Explosions

A supernova explosion deposits a kinetic energy typically $\sim 2 \times 10^{51}$ ergs into the interstellar medium. In the case of massive stars, the explosion is driven by the outward pressure exerted by escaping neutrinos produced in the collapse of the stellar core to a neutron star or a black hole. This is remarkable when one considers that the opacity of ordinary matter to neutrinos is so low that they can pass through a thickness of several light years of lead before being stopped! In the case of low mass stars ($M < 12M_\odot$ initially, and $M \sim 1.4M_\odot$ at the time of the explosion), the explosion results from an explosive thermonuclear burning of the stellar material under initially electron-degenerate conditions, which results in much of the star being burnt to a state of nuclear statistical equilibrium, which is dominated by the iron-peak elements. For a detailed discussion of these processes, see Arnett (1996).

Initially, in the fireball stage, the ejected material cools by adiabatic expansion following emergence of the shock at the stellar surface, and the bright optical display is driven by the diffusion of radiation from the core produced by the radioactive $\beta-$decay sequence ^{56}Ni $\rightarrow$ ^{56}Co $\rightarrow$ ^{56}Fe. As the stored radioactive energy, the ionised fireball recombines until finally it becomes optically-thin to the escape of radiation. At this time, it has become clumpy through the action of thermal instabilities and incomplete mixing, and, because the internal pressure is nearly equalised, moves out with a velocity which conforms closely to the law $v(r) \propto r$.

When the supernova ejecta first interacts with the surrounding interstellar gas with density ρ_o, at radius R, it initially drives a shock at a velocity v_s determined by the fastest moving ejecta with density $\rho(R,t)$ and with velocity $v(R,t)$; $\rho_o v_s^2 = \rho(R,t) v^2(R,t)$. Very shortly thereafter however, the blast wave is slowed down as interstellar material is swept into a shell. At this point a *reverse shock* is propagated back into the freely expanding ejecta, converting the kinetic energy of expansion into thermal energy. The hot shocked ejecta and the hot shocked interstellar material are separated by

a. Supernova Remnant: pre Sedov-Taylor Phase

b. Supernova Remnant: Sedov-Taylor Phase

c. Supernova Remnant: post Sedov-Taylor Phase

d. Stellar Wind Bubble
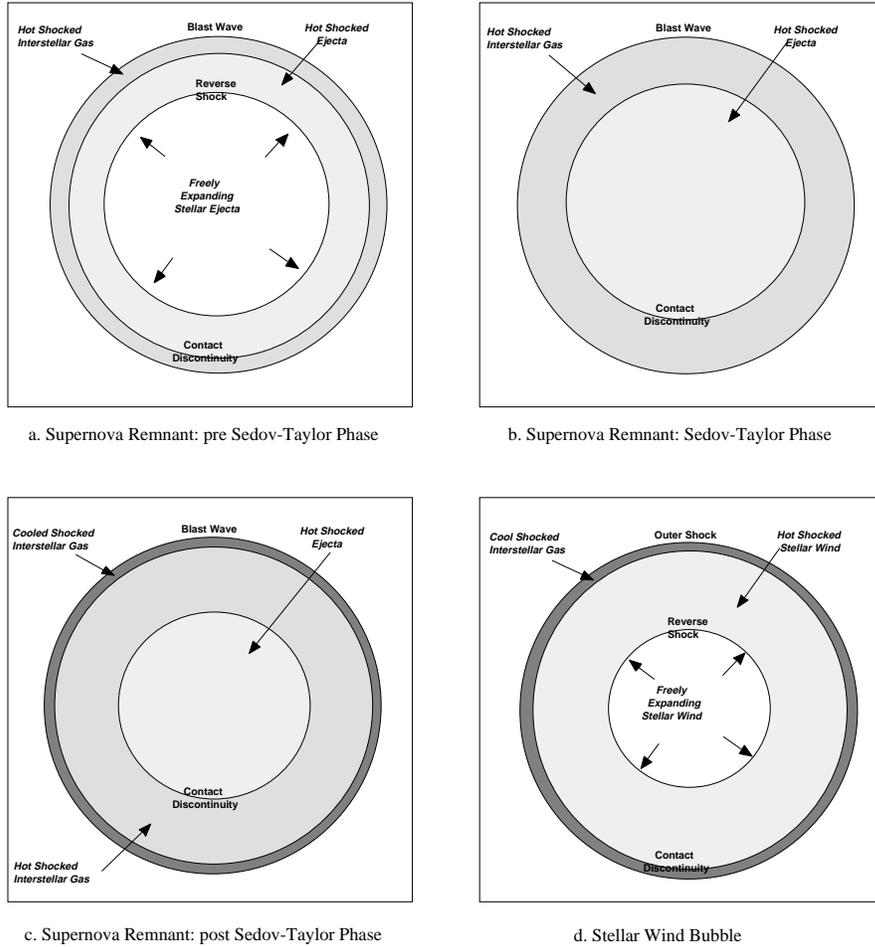
Fig. 8.4. The schematic hydrodynamic structure of a supernova remnant in the various evolutionary phases described in the text, and (d), of a stellar wind-blown bubble.

a *contact discontinuity*. The schematic hydrodynamic structure in this phase, and in subsequent phases is illustrated in Figure (8.4).

Usually the gas on either side of the contact discontinuity has different temperatures and densities, although the thermal pressures are the same. If the shocked stellar ejecta are cooler and denser than the swept-up interstellar gas, the contact discontinuity is *Rayleigh-Taylor* unstable, since the velocity of expansion is slowing with time. This configuration will produce mixing across the contact discontinuity.

Consider the case that the ejecta are not clumpy and the interstellar gas has a density ρ_0. When the mass of interstellar gas that has been swept up is quite a bit larger than the mass ejected, the reverse shock has swept back down to the explosion centre and all the ejecta have been shocked to high temperature. This marks the start of *Sedov-Taylor* phase of evolution. In this phase, we can regard the expanding bubble of hot plasma as suppling the piston to drive the blast wave outward. The specific internal (thermal) and the kinetic energies behind the strong adiabatic blast-wave shock at radius R are given by equations (8.24) and (8.25) respectively;

$$\epsilon_{int} = \epsilon_{kin} = \frac{9}{32}\,\dot{R}^2 \tag{8.34}$$

where $v_s = dR/dt = \dot{R}$. Since the blast wave is slowing down with time, the specific energy of the internal gas varies with radius within the bubble of hot gas. However, it turns out that this quantity, and the other thermodynamic and flow variables, vary in a self-similar way with respect to the dimensionless radial space variable r/R. The total energy in the bubble of hot gas, equal to the energy injected by the supernova explosion, E_0 since the radiative losses are negligible, is therefore given by:

$$\begin{aligned} E_0 &= \phi\frac{4\pi}{3}R^3\rho_0\left(\epsilon_{int} + \epsilon_{kin}\right) \\ &= \phi\frac{3\pi}{4}\rho_0 R^3\,\dot{R}^2 \end{aligned} \tag{8.35}$$

where ϕ is a structure parameter, a numerical factor of order unity which accounts for the distribution of specific energy within the bubble. Equation (8.35) represents the equation of motion of the bubble. When $t \to 0$, $R \to 0$, so that this equation has the solution:

$$R = \left(\frac{25}{3\pi\phi}\right)^{1/5}\left(\frac{E_0}{\rho_0}\right)^{1/5}t^{2/5} \tag{8.36}$$

and the instantaneous blast-wave velocity in the Sedov - Taylor phase is therefore:

$$v_s = \frac{2}{5}\left(\frac{25}{3\pi\phi}\right)^{1/5}\left(\frac{E_0}{\rho_0}\right)^{1/5}t^{-3/5} \tag{8.37}$$

The Sedov-Taylor phase is terminated when radiative losses behind the outward moving blast wave start to become important and a thin shell of cooled gas forms near this outer boundary. This phase sets in when the cooling timescale of the shocked plasma at the blast wave, found from equations (7.8), (8.37) and (8.26) becomes shorter than the dynamical expansion time, $\tau_{\exp} = R/v_s = 5t/2$; from equations (8.36) and (8.37). From this point onward, the *PdV* work done on the interstellar gas is subsequently radiated away, decreasing the total energy content of the hot bubble. The energy equation is therefore:

$$\dot{E} = -P4\pi R^2 \, \dot{R} \tag{8.38}$$

and the adiabatic equation of state gives:

$$E = \frac{4\pi}{3\,(\gamma - 1)} R^3 P \tag{8.39}$$

where E is the instantaneous energy content of the bubble. When combined with the equations of mass and momentum conservation:

$$M \qquad = \frac{4\pi}{3} R^3 \rho_0 \tag{8.40}$$

$$\frac{d(M \, \dot{R})}{dt} = 4\pi R^2 P \tag{8.41}$$

these equations provide a solution for the intermediate evolution of the shell; $R \propto t^{2/7}$ (McKee & Ostriker, 1977).

In the late phase of evolution, the stored thermal energy has been entirely radiated away, and only the momentum of the dense shell keeps the remnant expanding into the interstellar medium. This momentum conserving *snow-plough* phase of evolution was first considered by Oort (1946). The equation of conservation of momentum is, ignoring the momentum carried by the supernova ejecta at late times:

$$M_0 v_0 = \frac{4\pi}{3} R^3 \rho_0 \, \dot{R} \tag{8.42}$$

where M_0 is the mass of the ejecta, thrown out in the supernova explosion at a mean velocity v_0. This equation has the solution:

$$R = \left(\frac{3M_0 v_0}{\pi \rho_0} \right)^{1/4} t^{1/4} \tag{8.43}$$

Towards the end of this phase, the expansion velocity becomes sonic or subsonic with respect to the interstellar sonic or magnetosonic speed, and the remaining energy of the ancient supernova remnant is finally dissipated through incompressible turbulent cascade.

In summary, evolving in a constant density medium, the supernova expansion velocity is constantly slowed by its interaction with the surrounding medium, as we would expect intuitively. Initially, $R \propto t$, but later $R \propto t^{3/5}$, then $R \propto t^{2/7}$, and finally $R \propto t^{1/4}$. The middle two phases correspond to the phases where energetic radiative or partially-radiative shock waves are seen, first in the X-rays, but later at optical wavelengths as the stored internal energy is drained away in doing PdV work on the interstellar gas.

If the ejecta or the interstellar medium is cloudy, the evolution is more complex. In the case of cloudy ejecta, the onset of the Sedov-Taylor phase is blurred, as first the intercloud medium is thermalised, and then the clouds are thermalised over an extended period of time.

In the case of a cloudy interstellar medium, the blast wave sweeps over and compresses the denser regions, which then evaporate slowly into the shock-heated low density post-shock medium through the processes of turbulent shredding, mixing and thermal conduction (Cowie & McKee, 1977; McKee & Cowie, 1977); *see* Sections (7.3) and (7.4).

8.3.2 Stellar Wind Bubbles

All hot stars, be they either young and massive, or the old low mass objects at the cores of planetary nebulae, produce fast winds driven by radiation pressure. The physics of such winds is now well understood *eg* Sellmaier *et al.* (1993). In brief, radiation from the central star is scattered by the atmosphere, so that the momentum carried in the radiation field, L_*/c, is transferred to the atmospheric gas, to produce an outflowing wind. If v_w is the terminal velocity attained by the wind, and $\dot{M}_w$ its mass flux (typically of order 10^{-5} $M_\odot$ yr^{-1} for massive stars and 10^{-8} $M_\odot$ yr^{-1} for the central stars of planetary nebulae), then:

$$\dot{M}_w v_w = \frac{\eta L_*}{c} \tag{8.44}$$

Here, the factor η accounts for the fact that each photon may be scattered many times before it escapes, enhancing the total amount of momentum that can be deposited. This process is limited, since the wind cannot carry away more energy than is produced by the star:

$$\frac{1}{2}\dot{M}_w v_w^2 < L_* \tag{8.45}$$

Thus, from equations (8.44) and (8.45), $\eta < 2c/v_w$. In stars, the full application of the radiatively-driven wind theory shows that $\eta \sim 3 - 4$, typically. In such radiatively-driven winds, the outflow velocity is a factor $1 \lesssim \epsilon \lesssim 3$ times the escape velocity at the base of the outflow:

$$v_w = \epsilon \left(\frac{GM_*}{r_*}\right)^{1/2} \sim 1000 - 4000 \quad \text{km.s}^{-1} \tag{8.46}$$

This wind flows out in free expansion until it is affected by the interaction with the surrounding interstellar medium with density ρ_0. At this point it passes through an adiabatic shock, at an inner radius R_{in}, analogous to the reverse shock in the early supernova evolution. It is thermalised and feeds into a thick and hot pad of gas. In essence, it is this hot gas which is providing the piston to inflate the stellar wind bubble; *see* Figure(8.4d). The pressure P throughout the region between the inner shock and the outer shock can be taken as approximately constant, since the hot gas has a sound speed of order 500 km s^{-1}, and therefore the sound-crossing timescale in the hot plasma is much shorter than the dynamical expansion timescale of the bubble, which

has an expansion velocity in the range 20-100 km s^{-1}, typically. The pressure in the hot plasma is given by the rate of change of momentum per unit area of the stellar wind across the inner shock:

$$P = \frac{3\dot{M}_w v_w}{16\pi R_{in}^2} \tag{8.47}$$

The relatively low expansion velocity of the bubble will ensure that the outer shock is radiative, and isothermal at the temperature (~ 10000 K) of the pre-shock gas, which is kept ionised by the photons from the central star. We can therefore assume that the shocked interstellar gas forms a thin shell at the outer radius of the bubble. The equation of conservation of momentum is therefore:

$$\frac{d}{dt}\left(\frac{4}{3\pi}\rho_0 R^3 \dot{R}\right) = 4\pi R^2 P \tag{8.48}$$

or, equivalently,

$$\frac{P}{\rho_0} = \dot{R}^2 + \frac{1}{3}\ddot{R} R \tag{8.49}$$

Assume that the hot gas occupies a constant fraction $\phi = 1 - (R_{in}/R)^3$ of the total volume. The energy input by the stellar wind, $\dot{E}_w = \dot{M}_w v_w^2/2$ is equal to the sum of the rate of change of thermal energy in the hot gas and the rate of PdV work it does on the interstellar gas. This gives the energy conservation equation;

$$\dot{E}_w = \frac{d}{dt}(\frac{3P}{2}\frac{3\pi\phi}{4}R^3) + P\frac{d}{dt}\left(\frac{3\pi}{4}R^3\right) \tag{8.50}$$

The variable of pressure can then be eliminated from (8.49) and (8.50) to give the equation of motion of the shell. If we assume that the solution of this is a power law:

$$R = \Theta t^\beta \tag{8.51}$$

which has the correct boundary conditions, $R \to 0$ as $t \to 0$, then it follows by substituting for R and its derivatives with repect to time into the equation of motion that $\beta = 3/5$ and that:

$$\Theta = \left(\frac{125}{154\pi\phi}\right)^{1/5}\left(\frac{\dot{E}_w}{\rho_0}\right)^{1/5} \tag{8.52}$$

so that

$$R = \left(\frac{125}{154\pi\phi}\right)^{1/5}\left(\frac{\dot{E}_w}{\rho_0}\right)^{1/5} t^{3/5} \tag{8.53}$$

This solution was worked out by Castor, McCray & Weaver (1975) and Weaver *et al.* (1977). For further reading and for much more detail on the observation and the theory of stellar wind-blown bubbles, *see* Lozinskaya (1992).

8.3.3 Galactic Jet-Driven Bubbles

Advanced Section

In active galaxies, an accretion disk of gas forms around the central massive black hole. Some of this material, accreted from the innermost stable orbit, and falls through the event horizon while providing the energy for the luminous display. A small fraction of the inflowing matter is ejected in the form of a pair of well-collimated relativistic jets of material shot out in the polar directions. These jets may carry a mechanical energy flux which is an appreciable fraction ($\sim 30\%$) of the total luminosity. This material moves out, disturbed only by relatively weak internal shocks, until it interacts with the surrounding galactic (or intergalactic) medium. At this point, the jet flow is disrupted as it passes through a strong *termination shock* and the resultant hot shocked gas flows away from the shocked region, mixing in some of the shocked and swept-up galactic medium as it does so. However, we know from the observational and theoretical studies of jets on the kiloparsec scale, that these jets jitter about inside the cavity they have produced in the "dentist's drill" fashion first envisaged by Scheuer (1982). As a result, over time, the jet forms a pair of elongated bubbles of shocked and relativistic plasma which form a cocoon about the jets, and which provide the pressure to drive the shocks into the surrounding galactic medium.

The jet and its cocoon of shocked gas may display very complex hydrodynamics, depending on whether the jet is denser, or less dense than the surrounding interstellar gas, and upon whether the jet has a constant or variable energy flux. The system of shocks that may be set up in a generic jet (either from an active galaxy or from a young low-mass star) is shown in Figure (8.5).

As was shown by Begelman (1996), and Bicknell, Dopita & O'Dea (1997) the main factors which determine the evolution of the jet bubble are the jet parameters; its energy flux, $\dot{E}_J \sim 10^{43} - 10^{45}$ ergs s^{-1}, and the relativistic $\beta = v/c \sim 0.5 - 1$. Other parameters are the average pressure in the cocoon $P \sim 10^{-6}$ dynes cm^{-2}, and the density of the surrounding medium ($0.1 - 100$ cm^{-3}). Since in general this varies as with distance from the black hole, we write the density as $\rho(r)$. Towards the head of the cocoon, (in the x direction) the mean pressure is higher, because this is the area, A, over which the termination shock is jittering. Here we can take the pressure to be a factor ζ times the average lobe pressure where $\zeta \approx 2 - 10$.

With these definitions, the approximate equations of motion for the advance of the tip of the cocoon, and for its sideways expansion are from Begelman (1996):

$$\frac{dx}{dt} \approx \left(\frac{\beta \, \dot{E}_J}{\rho(x) \, cA} \right)^{1/2} \approx \zeta^{1/2} \left(\frac{P}{\rho(x)} \right)^{1/2} \tag{8.54}$$
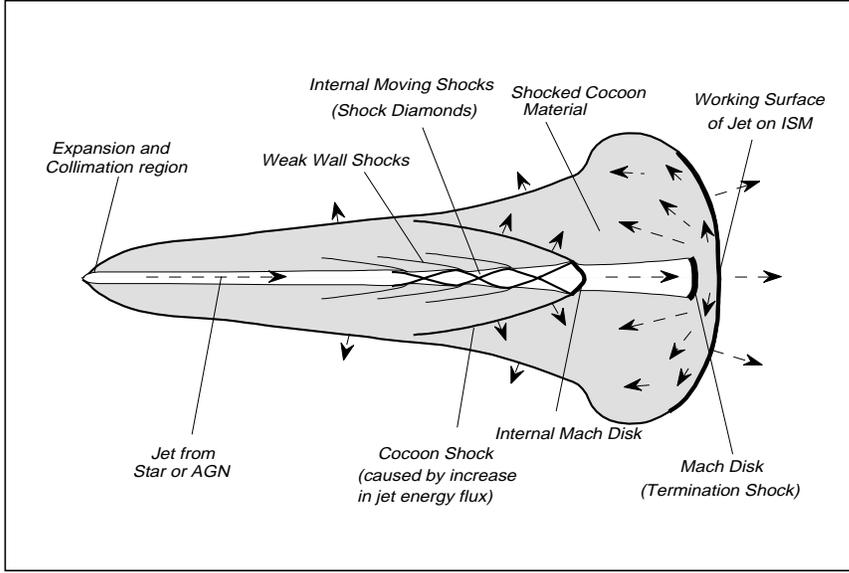
Fig. 8.5. The complex hydrodynamics of a jet powered by an active galactic nucleus, or a young star. The shocks interior to the jet are moving into a medium already in outflow, and may be oblique. As a result they tend to be weak, and show large proper motions. The termination shock which disrupts the jet is strong, and moves out slowly. The cocoon is a mixture of swept-up interstellar gas and of jet gas which has back-filled from the termination shock region.

$$\frac{dr}{dt} \approx \left(\frac{P}{\rho(r)}\right)^{1/2} \tag{8.55}$$

and, with V as the cocoon volume, the energy equation for the cocoon is:

$$\dot{E}_J = E + P\,\dot{V} \tag{8.56}$$

where $E = PV/(\gamma - 1) = 3PV$ (since we are dealing with a predominantly relativistic plasma) and $V \sim 2\pi x r^2/3$. If we assume a power-law distribution of the surrounding galactic medium; $\rho(r) = \rho_0(r/r_0)^{-\delta}$, and the boundary condition $x \to 0$ and $r \to 0$ as $t \to 0$ then these equations have the solution:

$$x = r_0 \xi^{1/(5-\delta)} \tag{8.57}$$

$$P = P_0\,\xi^{(2-\delta)/(5-\delta)} \tag{8.58}$$

where:

$$\xi(t) = \frac{(5-\delta)^3\,\zeta^2}{18\pi(8-\delta)}\left(\frac{\dot{E}_J\,t^3}{\rho_0 r_0^5}\right) \tag{8.59}$$

$$P_0(t) = \frac{9}{\zeta\,(5-\delta)^2}\,\rho_0\left(\frac{r_0}{t}\right)^2 \tag{8.60}$$

For $\delta = 2$ the bubble expands at a constant velocity. Expressing ρ_0 and F_E in terms of ξ and P_0 gives:

$$\dot{E}_J = \frac{2\pi(8-\delta)}{(5-\delta)}\frac{P_0 r_0^3}{t}\xi \tag{8.61}$$

$$\rho_0 = \frac{\zeta(5-\delta)^2}{9} P_0 \left(\frac{r_0}{t}\right)^{-2} \tag{8.62}$$

Typical dynamical ages are $\sim 10^6 - 10^7 \mathrm{yr}$.

The above solution implies that the shock velocity at the sides of the cocoon is:

$$v_{\mathrm{s}} = V_0 \left(\frac{r}{r_0}\right)^{\frac{(\delta-2)}{3}} \tag{8.63}$$

where the reference velocity of expansion,

$$V_0 = \frac{3\zeta^{1/6}}{[18(8-\delta)\pi]^{1/3}} \left(\frac{\dot{E}_J}{\rho_o r_0^2}\right)^{1/3} \tag{8.64}$$

is of order $(1-2) \times 10^3$ km s^{-1}.

Finally, the work done by the expanding cocoon on the ambient medium is:

$$P\frac{dV}{dt} = \frac{3}{8-\delta} \dot{E}_J \tag{8.65}$$

For $\delta = 2$ this amounts to $0.5\ \dot{E}_J$. If the cocoon is surrounded by radiative shocks this will be equal to the total shock luminosity.

8.4 The Radiative Properties of J-shocks

8.4.1 Radiation Properties of Shock Fronts

Advanced Topic

Thus far, we have considered a shock as a discontinuous change in the flow variables, accompanied by a discontinuous change in the Maxwell-Boltzmann distributions of the electrons, neutrals and ions. However, in real shocks, even those that are fully ionised, there will be a very small, but finite, width to the shock front, determined by the stopping timescale of the fast ions, and the equilibration timescale between the ions and the electrons. Within this narrow zone, in general, $T_e \neq T_{ion}$. When the plasma is only partly ionised, a problem first considered by Chevalier & Raymond (1978), the situation is even more complex. In this case, the shock occurs first in the ions.
For an ion of mass m_i and charge Ze the shock thickness is only of order of the ion gyroradius, $r = m_i c v_s / ZeB$, which gives an effective ion stopping

timescale of $m_i c / ZeB$, or a few minutes. These shocked ions may ionise neutral species directly, or transfer their thermal energy to the electrons, and which in turn collisionally ionise the neutrals. In addition, charge exchange reactions may also be important, allowing the neutrals to be thermalised more rapidly. Finally, streaming instabilities can be important, producing more rapid mixing of the different species.

It should be evident from this brief overview that the details of the shock front stucture can become exceedingly complex. Furthermore, not only can we have $T_e \neq T_{ion} \neq T_{neut}$, but also the distribution functions can become sufficiently unlike a Maxwell-Boltzmann distribution that even the concept of a local temperature applied to a particular species in the plasma is no longer valid.

Let us now consider the characteristic timescales for the various processes. The electrons will lose their streaming motion and become fully thermalised at the post-shock temperature on the electron-electron equipartition timescale, τ_{eeq}. These timescales have already been discussed in section (7.3.1) in the context of electron conduction, with the result that:

$$\tau_{eeq} = \frac{3m_e^{1/2} \left(kT_e\right)^{3/2}}{4\pi^{1/2} n_e e^4 \ln \Lambda_{coul}} \sim 0.065 v_{100}^3 n_e^{-1} \text{years} \qquad (8.66)$$

where, to provide a quantitative order of magnitude estimate for the timescale, we have scaled the results to a 50% ionised pre-shock medium, and a shock velocity, v_{100} given in units of 100 kms^{-1}. In such a meduim, the cooling timescale of the plasma in the post-shock region is a few hundred years, so it is clear that the concept of a unique local electron temperature remains valid throughout the shock.

The corresponding timescale for the electron-proton collisions is rather similar, since the volume density of protons is similar to that of the electrons, and the timescale scales as $\tau \propto M^2 v_{coll}$ where M is the reduced mass of the colliding system and v_{coll} is the characteristic velocity of the collision (see section (7.3.1) for the justification of this). The reduced mass $M = m_1 m_2 / (m_1 + m_2)$ of the system of colliding particles is now $M \sim m_e$ rather than $M = m_e/2$, but the characteristic velocity of the collisions is halved, approximately, so that we have:

$$\tau_{epeq} \sim 2\tau_{eeq} \qquad (8.67)$$

where τ_{epeq} is the electron-proton equipartition timescale. In a similar way, we may scale the proton-proton collision timescale. Here the reduced mass is $M = m_p/2$. However in this case, since the shock occurs first in the ions, the initial proton-proton collision timescale, τ_{ppcoll}, is governed by the streaming motions of the shocked and unshocked ionic populations; $v_{coll} \sim 3v_s/4$, so that the time over which the streaming motions are lost through proton-proton collisions is (approximately):

$$\tau_{ppcoll} \sim (n_p \sigma v_{coll})^{-1}$$

$$\sim \frac{3^3 m_p^2 v_s^3}{4^4 \pi n_p e^4} \sim 53 v_{100}^3 n_p^{-1} \text{years} \qquad (8.68)$$

this is much longer than the electron-proton collision timescale. What occurs instead is that the shocked protons provide thermal energy to the electrons, which then collide with the streaming protons. Because the electrons are lighter the momentum exchange per collision is lower, but the collision rate is higher thanks to the larger thermal velocity of the electrons. The electrons are therefore capable of removing the energy of the streaming motions on a timescale that is shorter than the proton-proton collisional timescale by a factor of $(m_e/m_p)^{1/2}$, approximately, so that the effective electron-proton collisional timescale is:

$$\tau_{epcoll} \sim 1.2 v_{100}^3 n_p^{-1} \text{years} \qquad (8.69)$$

Electron-proton collisions thereafter serve to thermalise the proton energy distribution function, on a timescale τ_{epeq}, which is short compared with τ_{epcoll}.

What about the neutral atomic species? Without Coulomb interactions, the timescale of interaction can be very long, and the mean free path is given by the gas kinetic theory. For example, heavy atoms such as helium are initially unaffected by the shock, and arrive in the post-shock region thermally cold, but at a streaming velocity, v_{coll}, which is 3/4 of the shock velocity. In principle then, thermalisation could be accomplished through helium-hydrogen collisions on a timescale of roughly:

$$\tau_{H,He} \sim (4 n_H \sigma v_{coll})^{-1} = \frac{4 m_H^{1/2}}{3^{1/2} \pi a_o^2 (kT_s)^{1/2} n_H}$$

$$\sim 170 v_{100}^{-1/2} n_H^{-1} \text{years} \qquad (8.70)$$

where n_H is the density of hydrogen atoms in the pre-shock gas. This timescale is much longer than the characteristic timescale of charge exchange reactions at high temperature; $\tau_{ce} = (R_{ce} n_{A,i})^{-1}$ which is typically of order 1 to 10 years in the example we are using. Thus charge exchange is more effective in producing thermalisation of the neutral stream. Even if charge exchange is ineffective, then the ionisation by electron impact, which occurs on a timescale $\tau_{coll} = (R_{coll} n_e)^{-1}$ provides for much more rapid ionisation of the stream of neutrals than the time over which the atom-atom collisions operate. In this way, both collisional ionisation and charge exchange processes serve to ionise the fast-moving atomic stream, which is then thermalised to the post-shock temperature by the strong Coulomb interactions.

Now let us consider how all of this affects the transient spectrum produced in the immediate post-shock region as the ionic and electronic components relax towards local equipartition of energy. In the shock front, itself, equation
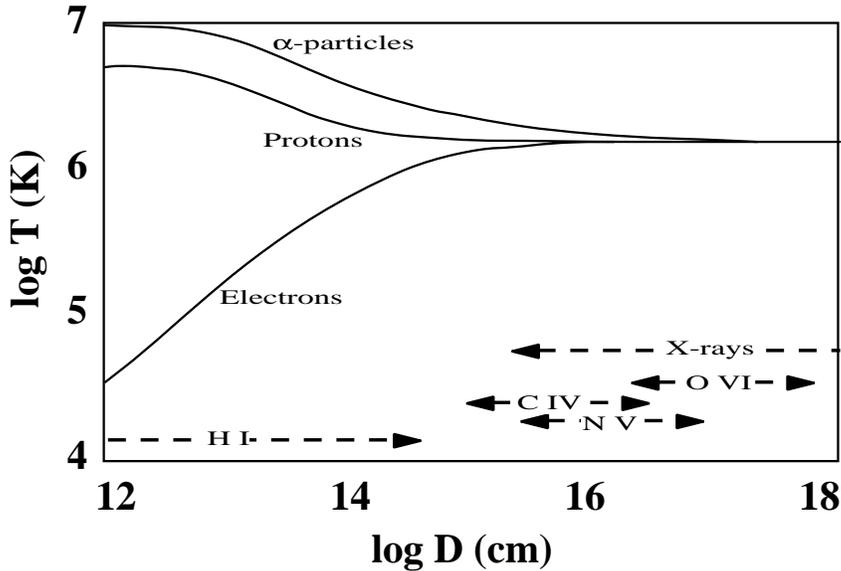
Fig. 8.6. Schematic structure of a fast shock front moving into a partially-ionised medium. Here D is the distance from the ionic shock. Ions and electrons come into equilibrium with one another through their Coulomb interactions at the same time as the plasma is becoming more highly ionised by collisional ionisations caused by both electron and ion impact.

(8.26) shows that the ions of different atomic weight may have quite different thermal temperatures. This equation also shows that, in principle, the electrons may have a temperature which is only (m_e/m_p) of the post-shock proton temperature. In practice, plasma instabilities will heat the electron gas to a higher temperature than this, but by an amount which is currently quite uncertain. As a consequence of the difference of electron and ion temperatures, excitation of electronic states in ions may occur much more readily by proton impact or alpha-particle impact rather than by electron impact, as is usually the case (Laming *et al.* 1996). Ionisation to progressively higher states occurs at a rate limited by the collisional ionisation timescale at the local temperatures. Schematically, therefore, the structure of the resulting shock front is as shown in Fig (8.6).

The emission spectrum of hydrogen is particularly interesting. When the electron temperature is lower than about 10^5K, the rate of the resonance charge exhange reaction, $H^0 + H^+ \rightleftharpoons H^+ + H^0$ exceeds the collisional ionisation rate. Thus, we obtain two populations of neutral hydrogen, one of which belongs to the pre-shock stream, and is thermally cold, and the other which has been thermalised at the post-shock temperature. As a consequence, collisional excitation – which proceeds more rapidly than collisional ionisa-

tion – produces hydrogen emission with both a narrow (pre-shock) profile, and an underlying broad (post-shock) emission component. Since the ratio of the charge exchange reaction rare to the collisional ionisation and excitation rates is a strong function of temperature, the ratio of the two components can be used to investigate the post-shock temperature structure. In addition, the width of the broad component provides a direct estimate of the shock velocity. This, combined with proper-motion studies provides a direct means of measuring the distance of young galactic supernova remnants.

Consider now the heavy ions. Since the forbidden lines have an emissivity which scales as $T^{-1/2} \exp[-\Delta E/kT]$, then at high temperatures they become very weak relative to the Balmer lines. Such *Balmer-dominated* shocks have been observed in young supernova remnants both in our Galaxy (*i.e.* SN1006, Schweitzer & Lasker, 1978 or the Cygnus Loop, Hester, Raymond & Danielson 1986 and references therein) and in the LMC (Tuohy *et al.* 1982). However, these high temperatures, combined with a relatively low ionisation state also ensures that the UV resonance lines are strongly excited. In the far UV, excitation of C IV λ1550Å, N V λ1240,46Å and O VI λ1032,38Å by fast ion impact can be more important than excitation by electrons. However, for some species such as He II λ1640Å the reverse is true, giving an observational means to probe the degree of equipartition between the electrons and the ions.

Finally, when electron temperatures are high enough to provide X-ray emission, the long ionisation timescales associated with the hydrogen-like and helium-like atomic species ensures that collisional excitation of the Lyman-like resonance lines of these species become much stronger than would be expected in an equilibrium model. Examples of such lines include the O VII lines at 570-700 eV, the O VIII lines at 650-850 eV, the Ne IX lines at about 900 eV and the Ne X lines at about 1.0 keV. These lines can be used to probe the age of the shock (*i.e.* Winkler, Bromley & Canizares, 1983).

8.4.2 The Structure of Radiative J-Shocks

The structure and the emission spectrum of radiative J-shocks of moderate velocity ($\lesssim$ 200 kms^{-1}) and in plasma of solar composition is now well understood following the work of Cox (1972), Dopita (1976,77), Raymond (1979) and Shull & McKee (1979). Following the shock itself, in the *equipartition zone*, the various components of the plasma relax towards equiparation of energy so that a unique post-shock temperature can be defined for all the plasma elements. The timescale for this is usually short with respect to any of the other timescales relevant to the plasma.

In the *ionisation zone* which follows, the plasma ionisation state increases rapidly, since pre-shock gas in an initially low ionisation state has been plunged suddenly into a very hot substrate. The timescale governing this process is the collisional ionisation timescale given (approximately) by equation (7.13); (*see also* figure 7.3). However, for shocks velocities slower than a

few hundred km s^{-1} the plasma does not have time to reach collisional ionisation equilibrium (CIE) before radiative losses become important. Radiative losses behind the shock will become dominant within the cooling timescale given by equation (7.8).

In the next zone, *the cooling zone*, the plasma cools approximately isobarically, the density increases, and the cooling timescale becomes shorter and shorter. As a consequence the ionisation state of the gas is governed primarily by the recombination timescales of the various ions, so that the degree of ionisation is higher than would be expected if CIE were to apply.

As the hot plasma cools, it is also producing EUV photons which are capable of ionising hydrogen. If hot enough, it can also produce soft thermal X-rays. When the temperature has fallen to about 10^4 K, the opacity due to atomic hydrogen increases rapidly, and the plasma begins to absorb the EUV photons from the hotter post-shock gas. As a consequence, we find a *photoionised zone* where, for a while, the plasma is almost in photoionisation equilibrium. In this region, the electron temperature is about 5000 to 8000 K, and hydrogen is typically about half ionised. Because of the compression produced through the shock, the ratio of the number of photons passing into this region per hydrogen atom is usully quite low for shocks slower than about 200 km s^{-1}, so that the photoionised region is weak and not very spatially extended. The effects of photoionisation in shocks faster than this, and in plasmas with unusually high heavy-element abundances will be considered in Chapter 11.

Half of the photons produced in the cooling zone pass upstream, back to the shock front and on into the pre-shock gas where they are absorbed. This produces a *photoionised precursor* to the shock. For shock velocities faster than about 120 kms^{-1}, the hydrogen entering the shock is effectively fully ionised. Since the ionisation fraction of the material entering the shock is a strong determinant of the subsequent shock structure, shock models need to take this into account by iterating between the shock structure and the pre-shock ionisation that this shock structure produces until a set of self-consistent pre-ionisation conditions have been obtained.

Based on detailed shock models, the time taken for the plasma to cool from when it was shocked to the point where hydrogen is about 50% ionised is:

$$\tau_{cool} \sim 200 \frac{v_{100}^{4.4}}{Z n_0} \text{years} \qquad (8.71)$$

where v_{100} is the shock velocity in units of 100kms^{-1}, n_0 is the pre-shock density (cm^{-3}) and Z is the gas-phase abundance of the heavy elements relative to solar values. Note how strongly this timescale depends on shock velocity.

Eventually, as we pass further downstream in the shock structure, the EUV photons are all used up, and, in the *recombination zone*, the plasma is free to continue its cooling and recombination. The timescale is set by the
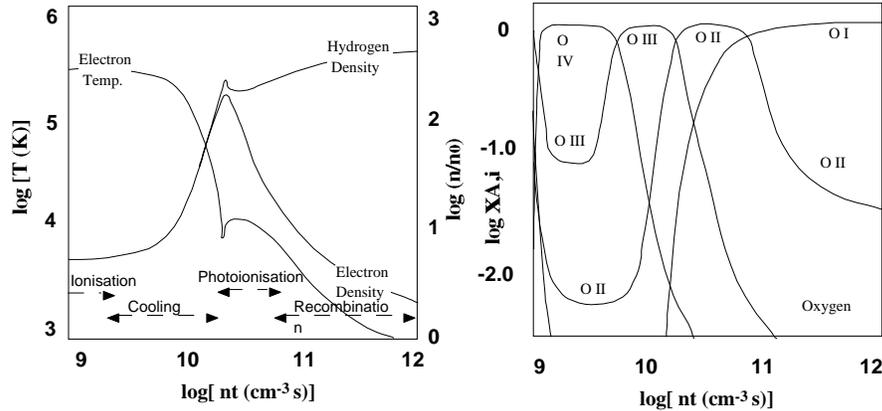
Fig. 8.7. The thermal and density structure (left) and the ionisation structure in oxygen ions (right) for a radiative shock of 150 km s^{-1}. The various zones described in the text are apparent. The tail of the shock is supported by magnetic pressure, which limits the compression. The ionisation of O II and O I are locked to hydrogen through charge-exchange reactions.

recombination timescale. This is easily estimated for hydrogen from equation (7.13) and from what follows equation (7.8);

$$\tau_{rec} \sim 8 \times 10^4 \frac{T_4^{0.72}}{n_e} \text{years} \tag{8.72}$$

years, where T_4 is the temperature in units of 10^4 K, and n_e is the electron density. This increases rapidly as the temperature and fractional ionisation decrease, so that the recombination tail can be quite extensive.

In the tail end of the recombination zone, when the temperature has fallen below about 3000K, the electrons have all but disappeared, and the temperature is too low to generate either near-IR or optical lines. As a consequence, the cooling timescale becomes very long, dominated by the spin-flip lines such as [C I]. If the plasma is dense enough, then these conditions favour molecule formation, so this zone may be regarded as a *molecular formation zone.*

In one of these phases, usually in the cooling phase, the pressure in the transverse magnetic field becomes comparable with the gas pressure. From that point, the shock becomes magnetically supported, limiting the compression factor that can be produced by the shock to that given by equation (8.31) or (8.32).

All these zones, with the exception of the first and last are shown in figure (8.7).

The spectrum of a radiative shock is very similar to that produced by a plasma cooling under non-CIE conditions. At the hot end of the cooling zone, the emission is mostly in lines and continua in the EUV region, which

are not directly observable from the ground. However, as the temperature cools below about 30000K, resonance lines of abundant species such as C II, C III, C IV, N V, O IV as a well as intercombination lines such as C III],N IV], O III] become strong in the 900-3000Å UV waveband. When the plasma is cooler than about 15000 K, the optical lines are emitted. Because the gas is mostly collisionally excited, even in the photoionisation or recombination zones, lines of low ionisation species are strong, especially those at the red end of the spectrum where the thermal energy of the electrons are comparable to the excitation energy of the atomic transitions. Such species include the [O I], [O II], and [S II] lines.

Throghout the whole shock structure each hydrogen atom undergoes at least one, and at most a few, recombinations. If the plasma is fully ionised on entering the shock, then the flux in the higher lines of the Balmer series increases as the mass flux through the shock, but is also somewhat dependent on the shock velocity. However, as the velocity increases, and the effect of photoionisation increases, and the Balmer line flux scales as the mass flux through the shock, and as the number of photons produced in the shock, which scales as the available enthalpy per unit mass divided by the mean photon energy. As a consequence, the Balmer flux increases somewhat more slowly than as the square of the shock velocity at low velocity, but this dependence steepens to become closer to the cube of the shock velocity at high velocities. This scaling can be used to estimate the pre-shock density if the shock velocity can be estimated.

In cooling between any two particular temperatures in the post-shock gas, there is only a certain amount of heat energy that can be radiated. As a consequence, the sum of the emission in the lines and continuum is determined by the application of the jump conditions between the two temperatures. We can therefore regard the various atomic and ionic coolants as being in competition with one another for their share of the total radiated energy. This has the consequence that, if we gradually increase the abundance of a particular coolant, the flux in any particular emission line of that species will at first increase in proportion to the abundance, but only for so long as the coolant contributes only a small fraction to the total cooling, *i.e.* it is a *trace coolant*. However, once the coolant has become the dominant one, the flux in its lines saturates at an absolute value related to the available enthalpy in the shock. Thus, trace coolants (especially in the UV) can be used to estimate the chemical abundances of the elements in the shock, but attempts to determine abundances from the strong lines are much less reliable (Dopita, 1976; 1977; Dopita *et al.* 1984).

8.4.3 Spectra of Old Supernova Remnants

Our theoretical treatment of the various evolutionary phases of a supernova remnant assumed that the remnant is expanding into a uniform interstellar

medium. The radiative phase, the phase at which the supernova remnant becomes bright at optical and UV wavelengths, was then assumed to commence with shell formation at the end of the Sedov-Taylor phase of evolution. However, the interstellar gas is always cloudy, and, for supernovae of Type II, the density contrast between the clouds and the inter-cloud medium may well have been enhanced when the clouds were swept up into the stellar mass-loss bubble of the precursor massive star.

When a cloud is swept up into a blast wave, a strong bow shock is formed in the upstream direction, and a lower velocity shock is driven into the cloud. It is this latter shock that first becomes radiative. The blast wave and the bow shock are usually too fast to become radiative within an evolutionary timescale, but these can be the source of X-ray and nonthermal radio emission from the remnant. The pressure on the leading edge of the cloud, the stagnation pressure, is somewhat higher than the mean pressure within the hot gas filling the interior of the supernova remnant. It is this pressure that drives the cloud shock. If ρ_c is the pre-shock density of the cloud medium, and ρ_0 is the pre-shock density of the inter-cloud interstellar medium then the relationship between the cloud shock velocity, v_c, and the blast-wave velocity, v_s, is given by:

$$P_{stag} = \rho_c v_c^2 = \xi \rho_0 v_s^2 \tag{8.73}$$

where ξ is a numerical factor in the range $1-4$, dependent on the geometry of the cloud. For optically radiative shocks observed in SNR, v_c is in the range $60 - 300$ km s^{-1}. Pre-shock cloud densities are usually less than 100 cm^{-3}. However, it is clear that densities may be even higher than this in dense molecular clouds, in which case the cloud shock is a slower molecular shock, $10 < v_c < 25$ km s^{-1}, bright in the infrared lines of H$_2$ and other species such as CO. Molecular cloud shocks are observed in a number of supernova remnants, amongst which IC443 and RCW103 are the best known examples.

For the cloud shock, the cooling timescale given by equation (8.71) provides a useful estimate of the time it takes to set up the steady-flow condition throughout the shock. If the shock is younger than the cooling time, then the shock will be only partially radiative. The hydrogen lines will be much weaker in such objects, as a result of the lack of time available for recombination, and in such shocks the UV lines and the lines of high-ionisation species must be stronger than in the steady-flow conditions. Such *incomplete* or *truncated* shocks have been used to explain the very large [O III]$\lambda\lambda$4959,5007/Hβ and [O III]$\lambda\lambda$4959,5007/[O II]$\lambda\lambda$3727,9 ratios observed in parts of the Cygnus loop supernova remnant (Fesen *et al.* 1982; Dopita & Binette 1983). In the Cygnus Loop, the pressure driving all the optically-visible shocks is similar; it is the pressure behind the blastwave. However, in this object the interstellar medium is particularly cloudy so that, depending upon the cloud or intercloud density, slow, fully radiative shocks ($n_c \sim 5 - 10$ cm^{-3}; $v_s \sim 100$ km s^{-1}), truncated shocks ($n_c \sim 1$ cm^{-3}; $v_s \sim 250$ km s^{-1}) and non-radiative shockfronts ($n_0 \sim 0.2$ cm^{-3}; $v_s \sim 400$ km s^{-1}) are all observed.

The luminosity per unit area of a fully radiative shock is given by equation (8.30). Since the shocks in supernova remnants are *hypersonic* (characterised by high Mach numbers), this can be written in terms of the pressure and density; $L_s \sim Pv_s/2$. Clearly therefore, optically lumionus supernove remnants occur where supernova explosions have occured in dense environments. Since the cooling time behind the shock is short in this case, such supernova remnants will also be compact and young. In the Galaxy, RCW103 and IC 443 are amongst the most luminous, and in the LMC the remnant N49 is the brightest.

The optical spectra of radiative supernova remnants are characterised by very strong forbidden lines of ions with low exitation potential, since these arise from collisional excitation in the partially ionised gas near the recombination zone of the shock. For example the [O II]$\lambda\lambda$3727,9Å lines are typically more than ten times the strength of the Hβ recombination line, and the [O I]$\lambda\lambda$6300,63Å and the [S II]$\lambda\lambda$6717,31Å lines are about as strong as Hα. These line ratios set supernova remnants apart from nebulae which are excited by UV photons from hot stars, and this difference has been very successfully used to identify supernova remnant candidates in the Magellanic Clouds (Mathewson *et al.* 1983) and beyond in local group galaxies (Sabbadin, D'Odorico, Dopita & Benvenuti, 1980; Blair & Kirshner, 1985). Proof that these objects are supernova remnants depends on obtaining optical spectroscopy showing large velocity dispersions, or on association of the optical object with a source of thermal X-ray or radio non-thermal emission.

8.4.4 Spectra of Herbig-Haro Objects

Advanced Topic

Estimates of the specific angular momentum of the clouds that become unstable to gravitational collapse and which subsequently form low mass stars always come to be out much larger than the angular momentum of such newly-formed stars. Left to itself, the collapsing gas cloud would form a flat rotating pancake of gas which contains most of the mass. The central star in this case would be small and rapidly rotating. In order to be able to form stars such as are observed, the material which is finally incorporated into the star must have exchanged its angular momentum with another portion of the gas, which gains so much angular momentum that it is constrained to flow out in an energetic wind. Such a wind becomes collimated above and below the mid-plane of the disk to form a pair of fast jets. These jets become visible as bipolar molecular ouflows. Where they drive shocks into the surrounding medium, or are shocked internally, they are visible at optical wavelengths as *Herbig-Haro (HH)* objects.

The most plausible model for this process has been developed in the x-wind model of Shu *et al.* (1994). In this, the role of the magnetic field is vital. If the star has mass M_* and a magnetic moment μ_* then the gas disk is terminated at an inner radius R_X;

$$R_X \sim \left(\frac{\mu_*^4}{GM_* \dot{M}^2} \right)^{1/7} \tag{8.74}$$

where $\dot{M}$ is the accretion rate of gas into this inner radius. The inner edge of this disk and the surface of the star are constrained to co-rotate, so that magnetic field lines connecting the inner edge of the disk (along which material must be accreted onto the star) are not wound up by differential rotation. The angular velocity, Ω_X, of the inner edge of the accretion disk is the Kelperian value, so that:

$$\Omega_* = \Omega_X = \left(\frac{GM_*}{R_X^3} \right)^{1/2} \tag{8.75}$$

To satisfy mass and angular momentum balance, the accreting material divides at R_X into a wind fraction $\dot{M}_W = f \, \dot{M}$ and a fraction that is funneled along the magnetic field lines leading down into the star, $\dot{M}_* = (1-f) \, \dot{M}$ where:

$$f = \frac{1 - \bar{J}_* + \mathcal{T}}{\bar{J}_W - \bar{J}_*} \tag{8.76}$$

where $\bar{J}_W$ and $\bar{J}_*$ are the specific angular momenta of these two streams in units of $R_X^2 \Omega_X$ and $\mathcal{T}$ is the viscous torque acting on the inner edge of the disk measured in units of $\dot{M} \, R_X^2 \, \Omega_X$. Since the specific angular momentum of the material flling on the star is so low, the wind carries a significant fraction of the mass, typically 25 - 35%.

At large distance from the pre-main sequence star, this wind is asymptotically collimated into a pair of oppositely-directed jets. Unlike the case of the jets in AGN where the jets are relativistic, carry a large energy flux and are usually less dense than their surroundings, these jets are cold and relatively dense, and are moving at typically 400 km s^{-1}. Nontheless, their hydrodynamics is similar to that shown in Figure (8.5).

Weak internal shocks in jets can be set up by self-excited reflective modes (Payne and Cohn, 1985). In this case, the shocked knots will form a string of Mach diamonds such as is shown in Figure (8.5). Their spacing, Δx, to jet diameter, D, ratio is related to the Mach Number in the jet and the density contrast between the interior of the jet and its surroundings:

$$\frac{\Delta x}{D} \leq (0.6 - 0.8) \frac{\mathcal{M}}{1 + (\rho_{int}/\rho_{ext})^{1/2}} \tag{8.77}$$

where the constant 0.8 applies to the first reflection pinching mode, a set of regular constrictions and expansions of the jet diameter, and the 0.6 applies to the helical Kelvin-Helmholtz pinching mode, a spiral twist in the jet surface like the surface of a rope. These internal shocks are seen whenever the external medium supplies a pressure which tends to confine the jet. When the jet

escapes into a low-density medium, it goes into free expansion, and becomes faint in the optical emission lines.

The internal jet shocks are highly oblique, and therefore weak. If θ is the angle between the pre-shock flow direction and the shock front, and v_J is the jet velocity, the transverse component of velocity, $v_\parallel = v_J \sin\theta$ is unchanged by passage through the shock, and the effective shock velocity is the perpendicular component of velocity, $v_{\perp,s} = v_J \cos\theta$.

In slow shocks the gas entering the shock front is only partially-ionised. The relatively high temperature in the post shock gas ensures a high rate of collisional excitation of the atomic hydrogen in the plasma. As a consequence, Lyα becomes the principal coolant in the plasma, and the electron temperature swiftly decreases following the shock. Decreasing electron temperatures increasingly favour collisional excitation of hydrogen over collisional ionisation, so an appreciable fraction of atomic hydrogen may persist throughout the shock structure when the shock velocity is less than about 100 km s^{-1}.

This has a number of effects on the spectra of jet shocks and weak wall shocks in HH objects. First, the Hα/Hβ ratio becomes large, thanks to the contribution of collisional excitation to the Hα line intensity. This is favoured by the existence of a resonance structure in the Hα excitation cross section just above threshold. Second, the flux of Lyα in the nebula becomes very large. This is not seen directly because of the very large optical depth in the Lyα line, but can be inferred indirectly in two ways, either through the strong enhancement of the hydrogenic two-photon continuum in such objects (*e.g* HH11, HH43, HH47; Dopita, Binette & Schwartz, 1982), or else through the molecular hydrogen fluorescence with the Lyα photons discussed in section (4.2.3), above. The final effect of partial ionisation of hydrogen in the pre-shock plasma is that charge-exchange at these high temperatures locks both the ionisation state of nitrogen and of oxygen to that of hydrogen. In addition, the very short cooling timescale in the post-shock gas depresses the degree of ionisation that can be attained in the shock. As a consequence, lines of low ionisation species such as [O I], [N I] and [S II] are unusually strong in these objects. In extreme cases, the [N I] $\lambda\lambda(5198+5200)$Å/ H$\beta$ ratio exceeds 3!

The termination shock is the point at which the organised outflow is disrupted due to the interaction with the interstellar medium. At this point, the jet passes through a Mach Disk, a strong shock, and is thermalised to high temperature. This hot plasma drives a bow-shock into the surrounding interstellar medium, which may also be characterised by a relatively high ($\sim$ 100 km s^{-1}) shock velocity. The PdV work done on the interstellar medium is supplied by the jet energy flux. However, since the area of the jet shock is smaller than the area of the bow shock, the jet shock is characterised by a higher pressure, and therefore a smaller cooling time. Thus, we can encounter cases such as HH47, where the jet termination shock is fully radiative, but the bow-shock is only partially radiative, and is therefore is characterised by

high [O III]λ5007Å/Hβ ratios, and high [O III] temperature, as measured by the $\lambda\lambda$4363/5007Å line ratio.

Many HH objects are characterised by very strong emission lines of refractory elements such as [Fe II], Ca II or Mg II. For these the gas-phase abundances seem to approach solar values. The most likely explanation for this is that the jet shocks arise in matter which originated very close to the star, and where the interstellar dust has been largely destroyed. In this case we should hope to distinguish spectroscopically between internal jet shocks and interstellar shocks, based on the ratios of the [Fe II] or Ca II lines with respect to Hβ. For a recent review of the spectroscopic data on HH objects, see Böhm & Goodson, 1997.

8.5 C-Shocks

So far we have considered shocks where, except for a thin region near the shock front itself, the flow variables in both the ionised and the neutral components of the plasma can be considered as changing discontinuously. For slow ($v_s \lesssim 50$km s^{-1}) shocks in weakly ionised molecular gas, the energy transfer between the ionised and neutral species may require a timescale which is longer than the characteristic cooling timescale of the plasma. Under these conditions, the hydrodynamic flow variables change continuously, and the shock is referred to as a *continuous* or *C- shock*, for which the theory was first constructed by Draine (1980).

In a C-shock, the shock occurs initially in the ions, which are in effect a trace constituent of the plasma. Since the magnetic field is pinned to the ionised part of the plasma, the magnetic field is compressed along with the ions. In the limit, the shock in the ions would be effectively an isothermal magnetically-supported J-shock with a compression factor given by equation (8.32). In practice, a discontinuity in the flow variables does not occur. This is because the Alfvén velocity in the neutral component of molecular clouds is typically about 1-3 km s^{-1} (Heiles *et al.* 1993), while the degree of ionisation that is produced by cosmic rays is only about 10^{-4}. The Alfvén velocity in the ion component is therefore (100-300) $\chi_{-4}^{-1/2}$km s^{-1}, where χ_{-4} is the fractional ionisation in units of 10^{-4}. This velocity may be lowered by a factor of as much as ten if there is a population of charged grains along with the ions, since these may contribute as much as 1% of the mass of the fluid. However, in most cases the Alfvén velocity in the charged component is comfortably greater than the C-shock velocity, so therefore the ionised component remains sub-Afvénic throughout the flow. In this case information about the approaching shock is communicated ahead of the shock in the form of a magnetic precursor (Draine, 1980), within which the flow variables change continuously. A shock in the ions will only occur if there exists a point in the flow at which the ion Alfvén velocity is less than the shock speed.

From the point at which the flow parameters in the ions begin to change, the neutrals will begin to drift with respect to the ions and the magnetic field. As a consequence there is a drag force produced by the collisional coupling between the ions and the neutrals; principally the ion-neutral elastic collisions and the charge-exchange process. At any point in the flow in which ion-neutral slippage occurs, there is a balance between the magnetic force on the ions and (ignoring the thermal pressure gradient) the collisional drag force on the neutrals. The first of these acts upstream in the direction of the magnetic field gradient while the drag term is determined by the drift velocity, and so acts in the downstream direction:

$$\frac{d}{dx}\left(\frac{B^2}{8\pi\rho}\right) = \langle\sigma v\rangle\, M_r n_i\, (v_n - v_i) \tag{8.78}$$

where n_n and n_i are the local density of the neutrals and ions, respectively, M_r is the reduced mass of the colliding species; $m_n m_i/(m_n + m_i)$ and $\langle\sigma v\rangle$ is the rate coefficient for momentum transfer. This is of order of the rate coefficient for resonant charge exchange in the orbiting approximation, given by equation (5.48); $\langle\sigma v\rangle \sim 1.5 \times 10^{-9}$ cm^{3}s^{-1}.

In a "pure" C-shock, the neutral fluid remains relatively cool, and therefore the flow (which will eventually become subsonic with respect to the sound speed) remains supersonic with respect to the magnetosonic velocity throughout, thanks to the field compression which has occured in the shock structure. If the heating becomes too great, then we can get a transition case in which the leading part of the shock has a C-shock structure, but we get a J-shock formed in the neutrals downstream.

8.5.1 The Structure & Spectrum of C-Shocks

Advanced Topic

The theory of the structure of C-shocks was first constructed by Draine (1980) and developed by Draine, Roberge & Dalgarno (1983). Notable contributions have been made by Chernoff (1987), Smith & Brand (1990) and by Smith in later papers (see Smith & MacLow 1997, and references therein). In addition the molecular chemistry of C-shocks has been developed by Flower *et al.* 1996).

An early success of the theory of C-shocks was that it provided a physical explanation for some observations (*i.e.* Nadeau & Geballe, 1979) which were otherwise inexplicable. These observations of molecular hydrogen in outflows in the vicinity of pre-main sequence stars showed supersonic wings in the line profiles extending out to ~ 100 km s^{-1}, which, along with excitation temperatures measured at about 2000K, suggested that the gas is being processed through shocks of quite high velocity. If these shocks were J-shocks, then the dissociation of molecular hydrogen should be complete, and such wings could not exist. The physical explanation of these observations is, of course, the low

temperature that is maintained by the neutral fluid throughout the C-shock structure.

Within the layer where there is slippage between the ions and the neutrals, the ion, electron and neutral species all have different temperatures. The ions are heated by the elastic collisions with the neutrals and cooled by inelastic collisions with the neutrals, and reach a characteristic temperature:

$$T_i = m_n \left(v_n - v_i\right)^2 / 3k. \qquad (8.79)$$

where, for a fully molecular plasma, $m_n \sim 7m_H/3$. This temperature is $\sim$30,000K in a C-shock with shock velocity, $v_s \sim 40$ km s^{-1}, assuming a drift velocity of roughly $v_s/2$. The electron temperature is lower than the ion temperature because the rate of momentum transfer through the elastic collisions is less effective.

For the neutrals, the heating per unit volume due to the ion-neutral scattering has to match the rate of radiative cooling per unit volume, $\dot{Q}\,(T_n)$, due to inelastic scattering at the characteristic temperature of the neutral species, T_n:

$$\dot{Q}\,(T_n) = \chi \left\langle \sigma v \right\rangle n_n^2 m_n \left(v_n - v_i\right)^2 \qquad (8.80)$$

The principal inelastic scattering processes are the excitation of vibrational and rotation transitions in H_2 or other abundant molecular species. Unlike the cooling functions discussed up to this point, the H_2 molecules are in a high density environment, and generally rather closer to LTE, so that the rate of radiative cooling per unit volume is proportional to density, rather than to density squared. The H_2 cooling rate is approximated by (Smith, 1993):

$$\dot{Q}\,(T_n)(T_n) \sim 4.2\mathrm{x}10^{-31} n_{H_2} T_n^{3.3} \mathrm{erg\ s}^{-1}\ \mathrm{cm}^{-3} \qquad (8.81)$$

These last two equations give the temperature of the neutrals $\sim$1750 K in a shock with $v_s \sim$40 km s^{-1}, taking the drift velocity to be roughly $v_s/2$, and also assuming $\chi =10^{-4}$ and $n_{H_2} = 10^4$ cm^{-3}. If the C-shock is in atomic gas, the cooling occurs mostly through far-IR electron spin flip transitions such as the [C II]$\lambda157.7\mu$mline.

Now, let us consider the flow in the shock in detail (Chernoff, 1987; Smith & Brand, 1990). Taking a steady, one-dimensional flow in the shock frame of reference, with shock velocity v_s, and ignoring the momentum carried by the ionised component itself, the laws of continuity and magnetic flux conservation, momentum and energy flux can be combined to give the momentum equations in the neutral and magnetic (ionised) fluids, respectively:

$$\frac{d}{dx}\left(m_n n_n v_n^2 + n_n k T_n\right) = F \qquad (8.82)$$

$$\frac{d}{dx}\left[\left(\frac{v_s}{v_i}\right)^2 \frac{B_0^2}{8\pi}\right] \qquad = -F \qquad (8.83)$$

and the energy equation for the neutral medium:

$$\frac{d}{dx}\left(\frac{m_n n_n v_n^3}{2} + \frac{\gamma}{\gamma - 1} n_n k T_n v_n\right) = F v_n + \Gamma + G \tag{8.84}$$

In these equations F is the drag force produced by the ion-neutral collisional coupling:

$$F = \langle \sigma v \rangle M_r n_n n_i (v_n - v_i) \tag{8.85}$$

Γ is the radiative loss rate term:

$$\Gamma = \Lambda(T_n) n_n^2 \tag{8.86}$$

and G is the rate of change of internal energy:

$$G = \frac{2\langle \sigma v \rangle M_r n_n n_i}{m_n + m_i}\left[\frac{1}{\gamma - 1}k(T_i - T_n) + \frac{m_i}{2}(v_n - v_i)^2\right] \tag{8.87}$$

It is helpful to make these equations dimensionless by first re-introducing the concept of the Alfvén Mach number of the flow:

$$\mathcal{M}_A^2 = \frac{v_s^2}{(B_0^2/4\pi m_n n_{0,n})} \tag{8.88}$$

where $m_n n_{0,n} = \rho_0$ is the density in the pre-shock flow (ignoring the mass of the ions). We also introduce the following (dimensionless) variables:

$$\zeta = \frac{v_n}{v_s} = \frac{n_{0,n}}{n_n}$$
$$\xi = \frac{v_i}{v_s} = \frac{n_{0,i}}{n_i}$$
$$t = \frac{kT_n}{m_n v_s^2}$$
$$\psi = \frac{m_n n_n \Lambda(T_n)}{M_r n_i \langle \sigma v \rangle kT_n} \tag{8.89}$$

and finally, we define a quantity with the dimensions of length, which is closely related to the characteristic length scale of the shock structure:

$$L = \frac{m_n v_s}{M_r \langle \sigma v \rangle n_{0,i}} \tag{8.90}$$

In terms of these variables, equations (8.82) to (8.84) reduce to:

$$L\frac{d}{dx}\left(\zeta + \frac{t}{\zeta}\right) = \frac{\xi - \zeta}{\xi\zeta} \tag{8.91}$$

$$L\frac{d\xi}{dx} = \frac{\xi^2(\xi - \zeta)}{\xi\zeta}\mathcal{M}_A^2 \tag{8.92}$$
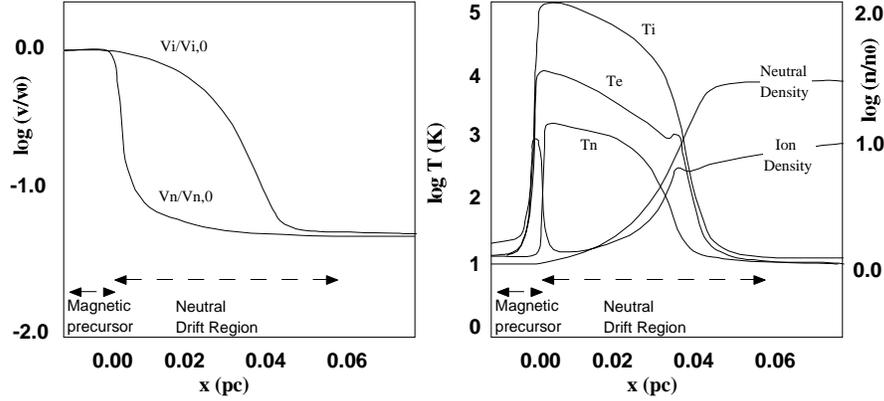
Fig. 8.8. The ion and neutral velocity profiles (left) and temperature and density profiles (right) for a steady-flow C-shock with $v_s = 40$ kms^{-1}, $B_0 = 100\mu G$ and intital density of 104 cm^{-3} (after Flower *et al.* 1996). The magnetic precursor and the neutral drift zones are approximately delineated.

and

$$L\frac{d}{dx}\left(\frac{\zeta^2}{2} + \frac{\gamma}{\gamma - 1}t\right) = \frac{\xi - \zeta}{\zeta} - \frac{\psi t}{\xi \zeta} \tag{8.93}$$

Eliminating x between these equations, and integrating the equation resulting from the combination of (8.92) and (8.93) with the boundary condition $\zeta = 1$ when $\xi = 1$ yields:

$$\frac{d\zeta}{d\xi} = \frac{(\gamma - 1)}{\mathcal{M}_A^2}\frac{[(\gamma\zeta/(\gamma - 1) - \xi)(\zeta - \xi) - \psi t]}{\zeta\xi^3(\zeta - \xi)(1 - \gamma t/\zeta^2)} \tag{8.94}$$

and

$$\zeta + \frac{t}{\zeta} = 1 + \frac{1}{2\mathcal{M}_A^2}\left(1 - \frac{1}{\xi^2}\right) \tag{8.95}$$

These equations can be integrated analytically in various limits such as energy-conserving flow or zero-temperature flow (Chernoff, 1987), or under more general and physically realistic limits (Smith & Brand, 1990). It is clear from the form of the equations that the nature of the particular shock solution is determined by the parameters $\mathcal{M}_A$ and ψ. This technique of reducing a hydrodynamic problem to one that can be solved analytically in terms of dimensionless parameters (frequently confined between zero and unity) is very powerful, and has been frequently used in obtaining particular solutions for hydrodynamical or, in this case, magnetohydrodynamical problems. Usually, however, they are solved numerically, with the analytic solution providing a test of the accuracy of the hydrodynamic code.

A typical solution for a steady-flow C-shock is shown in Figure (8.8). Initially, outflows from young stellar objects should drive shocks which are

J-type. It is only as the shock evolves, and the ratio of gas pressure to ram pressure falls, that the shock makes a transition towards the C-shock structure shown in Figure (8.8). This transition has been investigated using a full time-dependent magnetohydrodynamic treatment by Smith & Mac Low (1997). They find that rapid ion motions lead to the formation of an ion precursor, which then evolves towards a C-type structure as the J-shock in the neutrals decays and the flow relaxes over several Alfvén wave crossing timescales.

This evolution offers the potential to explain observations which had previously caused a good deal of confusion in their interpretation. For example, the molecular H_2 emission observed lines in the *Infrared Space Observatory* (ISO) 2-17μmwaveband come from a wide variety of rotational and vibrational states. This region includes the 0-0 S(1), the 0-0 S(7) the 3-2 S(1) and the 3-2 S(3) transitions. The excitation temperatures of these transitions range from about 1000 K up to over 19000 K, and so that their ratios may be used for the purpose of shock diagnostics. The usual way of doing this is to prepare a *column density ratio diagram*. This is a graph of the logarithm of the ratios of the column densities inferred from intensities of the various lines with respect to some reference line plotted against the excitation temperature of the various lines. The model fits are curves on this diagram, which ideally should pass through the points. When this was done, the observations seemed to be more indicative of J-shocks than of C-shocks, even though J-shocks needed were so fast that they would dissociate the molecular hydrogen. The answer to this problem suggested by the work of Smith & Mac Low (1997) is that the excitation is indeed due to C-shocks, but that these have not yet fully evolved to the steady-flow solution.

Notes on Chapter 8

- An excellent introduction to gas dynamics is in Dyson, J.E. & Williams, D.A. 1997 *The Physics of the Interstellar Medium, 2nd. Ed.*, Institute of Physics Publishing: Bristol, ISBN 0-7503-0460 (paperback). An elegant, but alas, somewhat difficult to obtain account of gas dynamics is by Kaplan, S.A. 1966, *Interstellar Gas Dynamics, 2nd Revised Edition*, ed. F.D. Kahn, Pergamon Press: Oxford.
- A classical description of shock physics is by McKee, C.F. 1987, in *Spectroscopy of Astrophysical Plasmas*, Eds. A. Dalgarno & D. Layzer, CUP:Cambridge, ISBN 0-521-26927 X (paperback), which has been very helpful in the preparation of this chapter. For those who are seeking a more advanced treatment, Draine, B.T. & McKee, C.F., 1993, Ann. Rev. A&A, 31, 373, is a good place to start.
- The physics of supernova explosions is dealt with very completely in Arnett, D. 1996, *Supernovae and Nucleosynthesis*, Princeton University Press: Princeton, ISBN 0-691-01148-6 (hardback), ISBN 0-691-01147-8 (paperback)

- For C-shocks, this research field is in a state of rapid evolution. However, a very clear exposition of the physics of such shocks is to be found in Smith & Brand (1990), and the recent time-dependent results of Smith & Mac Low (1997) is also worth reading.

Exercises

Exercise 8.5.1. Derive equations (8.30) and (8.32).

Exercise 8.5.2. A gas has a ratio of specific heats γ, and is at a pressure P and temperature T in a piston with an internal diaphragm. Suddenly the diaphragm is ruptured, and the gas expands into a vacuum down the cylinder.

Show that the maximum expansion velocity $v_{\exp}$ is, in terms of the initial sound speed of the gas, c_1; $v_{\exp} = c1[2/(\gamma - 1]^{1/2}$, and write down the temperature the gas achieves late in the expansion. **Hint:** write down the internal energy of the gas, and convert all of this into kinetic energy.

Exercise 8.5.3. In supernova remnant evolution, the Sedov-Taylor phase of evolution is terminated when the cooling timescale of the gas behind the blast-wave becomes comparable to the expansion timescale of the supernova remnant. Assuming that the supernova deposited 10^{51} ergs of energy, that the remnant is expanding into an interstellar gas with hydrogen atom density of 1.0 cm^{-3}, and that the cooling function Λ is constant with temperature and equals 10^{-22} erg cm^3 s^{-1}, compute the approximate radius at which the remnant becomes radiative. *Hint:* use equations (7.8), (8.26) , (8.36) and (8.37), and assume $\phi \sim 1$.

Exercise 8.5.4. a. Derive equations (8.52) and (8.53) from (8.49), (8.50) and (8.51). b. Estimate the ratio R_{in}/R.

Exercise 8.5.5. The nuclear-burning lifetime of massive stars is approximately $\tau = 5(M/40M_\odot)^{-0.4}$ Myr, and in this time thermonuclear reactions convert approximately 0.3% of the rest mass energy into radiative energy. Using: $M_\odot = 1.998 \times 10^{33}$ g and $R_\odot = 6.6 \times 10^{10}$ cm, and showing working,

a. What is the total luminous energy released by, and the luminosity during its lifetime of, a 60 $M_\odot$ star?

Answer: 3.23×10^{53}erg, 2.40×10^{39}erg s^{-1}

Stellar winds are driven by the momentum flux in the radiation field, and carry a momentum flux $\dot{p} = \eta L/c$, with $\eta \sim 3$, as a result of multiple scattering in the wind. The wind terminal velocity is given by $v_\infty = \epsilon (GM/r)^{1/2}$ where r is the radius of the star, and ϵ is a dimensionless factor, typically about 2.

b. Taking $\eta = 3$ and $\epsilon = 2$, and assuming a stellar radius of 20 times that of the sun, what is the momentum flux and stellar wind velocity in the stellar wind of our 60 $M_\odot$ star? **Answer:**2.41×10^{29}g cm s^{-1} and 1560 km s^{-1}

c. Calculate the ratio of mechanical energy flux to the luminous energy flux in this star during its lifetime.

At the end of its life, the star explodes as a supernova, and deposits a total of 2×10^{51} ergs into the surrounding interstellar medium.

d. What is larger, the energy deposited by the stellar winds, or the energy produced in the supernova explosion? **Answer:** The winds

e. The density of the surrounding interstellar medium is 10^{-23} g cm^{-3}. Using equation (8.53),what is the diameter of the mass-loss driven bubble (in parsecs, 3.08×10^{18} cm) when the supernova explodes?

Exercise 8.5.6. A protostar has been producing a highly supersonic and highly collimated jet of material, which has propagated a long way into the surrounding interstellar medium. The mass-loss rate $\dot{m}$ in the jet is constant, and the jet velocity v_0 is also constant. Suddenly, the mass-loss rate and the velocity of the jet start to increase linearly with time, until after time t_1 both are exactly twice their intial value. After some time, a dense sheet of material forms in the flow, bounded on both its inner and outer surfaces by shocks, which can be assumed to both radiative and isothermal. This dense layer is called a *working surface*.

a. Explain why such a surface must form, and calculate how far from the sar, and how long after the jet properties first start to change, that the working surface makes its appearance.

b. Assuming that the working surface can be approximated by a thin shell, compute an equation describing its equation of motion reckoned from the time and the place that it first appears.

c. Give an algebraic expression for the luminosity of the working surface as a function of time.

d. What would the equation of motion of the working surface be in the general case, where the velocity increase in the jet is a factor θ times its initial value and the mass loss rate increases by a factor ψ. Again, reckon the origin of coordinates to be the place and the time where the working surface first appears.

Hint: (a) Remember that two flows cannot overrun each other. In general they are separated by a *contact discontinuity* across which the pressure is balanced (even if the temperatures, molecular weights or densities are not).

(b) The relative kinetic energy of the colliding flows is radiated away (and this quantity gives the answer to part (c) of this question), so for this part, just worry about the conservation of mass and momentum.

Exercise 8.5.7. A powerful jet is propagating from the nucleus of an active galaxy with a relativistic $\beta = v/c$. At its head, gas is shocked at a working surface shock, thermalised, and back-flows to form a cocoon of hot plasma around the jet. The gas pressure in this cocoon pushes a transverse shock into the surrounding interstellar medium.

a. Assuming a constant density in the galaxian interstellar medium, ρ, a jet energy flux, $\dot{E}$, and a jet opening angle θ, give an expression for the velocity of advance of the jet as a function of radius in the galaxy, assuming

that the bow-shock in the interstellar medium always remains close to the jet shock. Compute this velocity for $\dot{E} = 10^{44}$ erg s^{-1}, $\beta = 0.95$, $\rho = 10^{-24}$ g cm^{-3} and $r = 3 \times 10^{21}$ cm (1.0 kpc).

Hint: Equate the ram pressure in the jet to the ram pressure across the shock.

b. Assuming that the hot gas pressure in the cocoon is constant, P, and that the transverse shocks are radiative, compute the transverse expansion velocity of the cocoon, v_0. What will this be (in km s^{-1}) for $P = 10^{-6}$ dynes cm^{-2} and $\rho = 10^{-24}$ g cm^{-3}?

c. Give an expression for the luminosity of the radiative shocks per unit area, and compute this for the above values of pressure and density.

Hint: Equate the ram pressure in the jet to the ram pressure across the shock and then compute the work done in the expansion.

d. At time t_0, the jet suddenly breaks out of the confining galactic interstellar medium, and the pressure in the cocoon drains away. Assuming a constant density in the galaxian interstellar medium, derive an expression for the velocity of the transverse shocks at time. Show your working.

Answer: $v(t) = v_0/[3(t/t_0) - 2]^{2/3}$

Exercise 8.5.8. A shock in a Herbig-Haro object is propagating through neutral gas at a velocity of 100 km s^{-1}. The gas has a hydrogen number density of 10 cm^{-3}, a temperature of 10,000K, and also contains 10% (by number) of helium atoms but negligible numbers of heavier atoms.

a. Assuming a monatomic gas, that magnetic fields are negligible, and that the molecular weight of neutral gas is $\mu = 1.4$, calculate the immediate post-shock temperature and the compression factor across the shock.

b. What would be the post-shock temperature if the gas entering the shock was ionized, assuming that the molecular weight of fully ionized gas is 0.70?

c. Assume that, following the shock in neutral gas, each hydrogen atom is excited to the $n = 2$ level (10.2 eV above ground) on average two times before it is ionized (which requires 13.6eV of energy) and that this is the only cooling process. If, in the post-shock plasma, the gas cools isobarically, and that other cooling processes are negligible, at what temperature does hydrogen becomes fully ionized? In this case, what is the total compression factor of the pre-shock gas at this point?

d. Assume that the gas has fully recombined before it cools to a temperature of 10,000K. What is the (exact) compression factor of the gas when the gas has cooled to 10,000K?

e. If the pre-shock gas is threaded by a purely transverse magnetic field of 30μG, what is the sound speed, the Alvén speed and the magnetosonic speed in this gas?

f. What would be the maximum compression factor across the shock in this case?

9. The Theory of Photoionised Regions

"How far that little candle throws his beams!
So shines a good deed in a naughty world"
— Shakespeare (The Merchant of Venice, Act5)

In this chapter, we move from considering plasmas that are excited by predominantly collisional means to those that are heated by energetic photons coming from a photoionising source, be it a star, a nucleus of an active galaxy, or whatever. Such photoionisation-dominated plasmas are very commonly observed in galaxies, and historically were one of the first to be studied in detail, thanks to their bright and prominent emission line spectra at optical wavelengths.

Within a photoionised region, the column density of neutral hydrogen has to be very small. The reason for this can be readily understood. In a normal plasma, hydrogen is by far the most abundant of the elements, outnumbering the helium atoms by about ten and all the heavier elements taken together by a factor of several hundred. As a result, the opacity of the ISM at EUV frequencies above the Lyman Limit, ν_0, is dominated by the absorptivity of hydrogen, derived from equation (5.37):

$$\kappa_\nu = \sigma_0 \left(\frac{\nu}{\nu_0}\right)^{-3.5} n_{H^0} \tag{9.1}$$

where n_{H^0} is the volume density of atomic hydrogen, and σ_0 is the absorption cross-section at the threshold frequency for photoionisation of hydrogen $(6.3 \times 10^{-18}$ cm^2). From the model of the H I distribution in the local solar neighbourhood, given by (7.63), the H I gas density at the mid-plane of our Galaxy near the sun is about 0.6 cm^{-3}. Thus for EUV photons at the threshold for ionisation, an optical depth of unity is reached after only 0.09 pc, and for EUV photons just below the threshold of ionisation of helium, this point is reached after 0.6 pc. However, H II regions in galaxies like our own are observed to range in size up to radii of order 100 pc. This can only be explained if within them, hydogen is of order 99% ionised, on average.

It is the absorption of the EUV photons by these trace amounts of neutral hydrogen that provides the local heating rate, $\dot{Q}$, to the ionised plasma. In equilibrium, this has to be balanced against the local cooling rate, provided

mostly by collisional excitation of forbidden emission lines in heavier atomic species. If $I(\nu)$ is the intensity of the local radiation field which extends up to a maximum frequency of ν_{max}, then the heat balance equation is:

$$\dot{Q} = \int_{\nu_0}^{\nu_{\mathrm{max}}} \frac{(\nu - \nu_0)}{\nu} I(\nu)\, \kappa_\nu d\nu = \Lambda n n_e \tag{9.2}$$

where Λ is the cooling function (7.1) defined for the particular local ionisation conditions and chemical abundances. The term $(\nu - \nu_0)/\nu$ in this equation arises because the electrons produced by photoionisation have to first overcome the ionisation potential of hydrogen to escape the atom.

Most H II regions tend to have an electron temperature in the vicinity of 10^4 K. This is because in this temperature region the cooling rate rises rapidly by several orders of magnitude with temperature, while the heating rate decreases slowly with temperature, since collisional ionisations of hydrogen tend to deplete the population of neutral hydrogen atoms, which in turn lowers κ_ν. As a consequence, there is a strong "thermostat" effect tending to keep the temperature constant.

The local radiation field which controls both the local ionisation state and the electron temperature is determined by the radiative transfer through the nebula. For a simple plane-parallel slab nebula, this is given by equation (4.9):

$$\frac{dI(\nu)}{ds} = -\kappa_\nu I(\nu) + j_\nu \tag{9.3}$$

Here, the local source term, j_ν, arises because recombination in hydrogen atoms may occur directly from the continuum back down to the ground state, producing a photon just above the Lyman limit, which is able to ionise hydrogen in its turn.

Equations (9.1) to (9.3) make apparent the intimate coupling between the radiative transfer, and the resulting ionisation and heat balance in an H II region. It is this coupling which provides the rich phenomenology of observed H II regions, and which we will investigate in this chapter.

9.1 Photoionisation of Hydrogen H II Regions

9.1.1 Equilibrium Ionisation & Ionisation Parameters

Let us consider, for simplicity, the photoionisation of a nebula composed of hydrogen only. When such a nebula is in photoionisation equilibrium, then at each point in the nebula, the rate of recombinations must match the rate of photoionisations. From (9.3) and (9.1), this equilibrium condition is (neglecting the local radiation source term):

$$\alpha\left(T_e\right)n_{H^+}n_e = \int_{\nu_0}^{\nu_{\mathrm{max}}} \kappa_\nu \frac{I(\nu)}{h\nu} dv$$

$$= n_{H^0} \int_{\nu_0}^{\nu_{\mathrm{max}}} \sigma_0 \left(\frac{\nu}{\nu_0}\right)^{-3.5} \frac{I(\nu)}{h\nu} dv \qquad (9.4)$$

where $\alpha\left(T_e\right)$ is the effective recombination rate for hydrogen. Let us first simplify the integral to $S\langle\sigma\rangle$, where S is the number of source photons passing through a unit volume locally $(\mathrm{cm^{-2}s^{-1}})$ and $\langle\sigma\rangle$ is the average photoionisation cross section $(\mathrm{cm^2})$, weighted according to the shape of the ionising source spectrum. If χ is the fractional ionisation of hydrogen, then we can also set $n_e = n_{H^+} = \chi n$, and $n_{H^0} = (1-\chi)n$ to give:

$$\frac{\chi^2}{1-\chi} = \frac{\langle\sigma\rangle}{\alpha\left(T_e\right)}\frac{S}{n} \qquad (9.5)$$

Since $\langle\sigma\rangle$ and $\alpha\left(T_e\right)$ usually vary only by a factor of a few, while S and n may take almost any value, this equation clearly shows that the local ionisation state of the plasma is primarily determined by the ratio S/n, which has the dimensions of a velocity. This is called the *ionisation parameter*, q and it has a simple physical meaning. Suppose that, instead of being incident on a plasma in photoionisation equilibrium, the photon field S was instead incident on atomic hydrogen of density n. In this case, the flux of photons would have to match the number of new ionisations they produce, so that the boundary of the ionised region will advance at a velocity dx/dt given by:

$$S = n\frac{dx}{dt} \qquad (9.6)$$

Thus q is simply the velocity of the *ionisation front* that the radiation field could drive through the medium. In addition to the ionisation parameter q, the *dimensionless ionisation parameter* $\mathcal{U}$ is also frequently used (confusingly this is also referred to as the ionisation parameter by those that use it). This is simply obtained by dividing q by the speed of light, c. It also has a simple physical interpretation since:

$$\mathcal{U} = \frac{q}{c} = \frac{U}{\langle h\nu\rangle n} \qquad (9.7)$$

where U is the local photon energy density. Thus $\mathcal{U}$ is simply the local ratio of the density of photons to the density of atoms.

Occasionally, there is used another type of dimensionless ionisation parameter, Ξ, which is physically the ratio of the radiation pressure to the gas pressure. This proves to be useful when considering the phase stability of the ISM. It is simply related to the ionisation parameter $\mathcal{U}$:

$$\Xi = \left(\frac{P_{rad}}{P_{gas}}\right) = \left(\frac{\langle h\nu \rangle \, S/c}{2nkT_e}\right) = \mathcal{U}\frac{\langle h\nu \rangle}{2kT_e} \tag{9.8}$$

With these definitions, let us return to equation (9.5). This shows that the plasma will be half-ionised when $q = \alpha\left(T_e\right)/2\left\langle \sigma \right\rangle$. Let us substitute numerical values. $\langle \sigma \rangle$ is taken from (9.1) with $\langle \nu \rangle \sim 1.5\nu_0$, and $\alpha\left(T_e\right) = 4.2 \times 10^{-13}$ cm^3 s^{-1}; derived from the equation following (5.18) with $T_e \sim 10^4$ K. Hydrogen becomes less than 50% ionised when q drops below about 10^5 cm s^{-1}, a figure which is (coincidentally) roughly equal to the sound speed in the neutral plasma. This point could be taken as the "edge" of the region of ionised gas. Defining $\langle \tau \rangle$ as the optical depth from this edge ($\chi = 1/2$), to another point within the ionised region, we have, from equation (9.5),

$$\frac{\chi^2}{1-\chi} = \frac{1}{2}\exp\left[\langle \tau \rangle\right] \tag{9.9}$$

with

$$\langle \tau \rangle = \int_0^x \langle \sigma \rangle \, n(1-\chi)dx = \int_0^\lambda (1-\chi)d\lambda \tag{9.10}$$

where we have introduced a dimensionless length scale $\lambda = \langle \sigma \rangle \, nx$. These equations provide a readily integrable approximate solution to the ionisation balance in a plane-parallel slab of ionised hydrogen, for any incident radiation field intensity. This is shown in Figure(**??**), and the local ionisation parameter in the flow is indicated. It is evident that for the values of q encountered in typical H II regions ($3 \times 10^7 \lesssim q \lesssim 3 \times 10^8$ cm s^{-1}), hydrogen remains more than 98% ionised throughout the bulk of the ionised region. The column of material (cm^{-2}) that can be kept ionised at any particular ionisation parameter and density is referred to as the *Strömgren Column*, in honour of the person who first developed the theory of H II regions (Strömgren, 1939).

In a spherical nebula centred on the exciting source of radiation, we can use the idea that hydrogen is effectively fully ionised throughout the ionised volume to estimate the equilibrium radius of the ionised region of radius R_s (the *Strömgren Radius*). This is obtained by setting the number of ionising photons equal to the number of recombinations within the ionised volume. If S_* is the number of ionising photons produced by the star, then:

$$S_* = \frac{4\pi}{3}\alpha\left(T_e\right)n_{H^+}n_e R_s^3 \tag{9.11}$$

With the assumption of a fully ionised plasma, $n_{H^+} = n_e = n$, we then have:

$$R_s = \left(\frac{3S_*}{4\pi\alpha\left(T_e\right)n^2}\right)^{1/3} \tag{9.12}$$

This equation shows that, the lower the density, the larger the H II region produced. However, this larger size does not make the H II region any easier
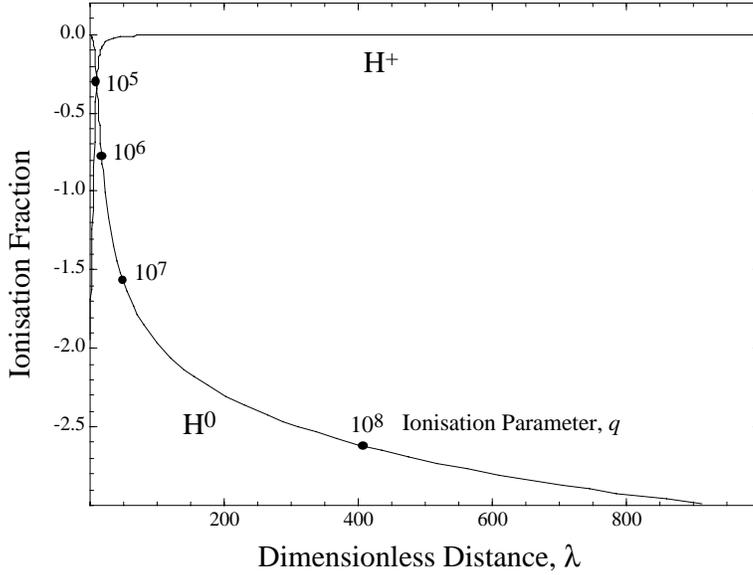
Fig. 9.1. The approximate ionisation structure of hydrogen in an H II region, obtained by solution of equations (9.9) and (9.10). The ionising source is to the right. The points at which the local ionisation parameter is 10^5, 10^6, 10^7 and 10^8 are marked.

to observe. The total flux from the nebula in a hydrogen recombination line is directly proportional to the total number of recombinations, which in turn, using equation (9.11), is simply proportional to the flux from the central ionising source, S_*, and is independent of the density. It is different for the surface brightness. The surface brightness of the nebula in a recombination line is proportional to the *emission measure* of the nebula, which from equation (1.3) is proportional to the product $n^2 R_s$. From equation (9.12), this is proportional to $n^{4/3}$. Thus, low density (large) nebulae have much lower surface brightnesses than high density (compact) nebulae of the same total flux, rendering them more difficult to observe against the night sky emission.

The ionisation balance within this spherical nebula is readily obtained from equation (9.5), by integrating from a radius $r = 0$, allowing for the spherical divergence of the radiation field, and its attenuation through the ionised volume (ignoring the local diffuse field):

$$\frac{\chi^2}{1-\chi} = \frac{\langle \sigma \rangle}{\alpha\left(T_e\right)} \frac{S_* \exp\left[-\tau\left(r\right)\right]}{4\pi r^2 n}$$
$$= \frac{\tau_s}{3} \exp\left[-\tau\left(r\right)\right] \left(\frac{r}{R_s}\right)^{-2} \tag{9.13}$$

where:

$$\tau\left(r\right) = \int\limits_{0}^{r} \left\langle\sigma\right\rangle n(1-\chi)dr \qquad (9.14)$$

Here, we have defined τ_s as the mean optical depth of an sphere of neutral hydrogen out to the radius of the Strömgren sphere; $\tau_s = \left\langle\sigma\right\rangle nR_s$. These equations are very similar to those of the plane parallel case, equations (9.9) and (9.10), and again, can be solved by numerical integration. However, in this case they are dependent upon the geometry. The intensity of the radiation field increases more quickly towards the source than the plane parallel case, thanks to the spherical divergence term, so that the ionisation state of the plasma near the inner edge of the nebula is higher.

Because of the very steep dependence of the hydrogen opacity with frequency, the ionising photons with energies close to the Lyman limit are absorbed first. Consequently, the radiation field of the central source is "hardened" (*i.e.* the average photon energy shifts towards higher frequencies) during its passage through the nebula. Thus, although the space density of photons is decreasing, the energy delivered to the electron gas per ionisation is increasing. Under certain geometries, this radiation hardening increases the energy input per photoionisation sufficiently to counteract the decrease in the product of the radiation density and the neutral fraction as we move outward in the nebula. This may lead to an increase in the electron temperature towards the outer boundary of the nebula, even though the state of ionisation continues to fall towards this boundary.

9.1.2 The Diffuse Field: Case A & Case B.

So far we have simply ignored the j_ν term in the equation of transfer (9.3). However, as has already been briefly mentioned, there is a local production of a diffuse ionising radiation field resulting from recombination of hydrogen directly to the ground state. Because of the selection rules, electrons which are captured during recombinations of hydrogen to the ground $^1S_{1/2}$ state must be those that were in a (continuum) virtual P state before the collision. The statistical probability of this decreases very quickly for energies above the Lyman limit, and therefore we can regard the spectrum of the diffuse photons as being a delta-function just above the Lyman limit.

Unlike the stellar photons, which are directed outwards, these diffuse photons may cross the nebula in any direction from the point of their production. The escape of these photons from a spherical region in which the optical depth at the Lyman limit, τ_{ν_0}, is therefore governed by an escape probability of the form of equation (4.16). According to equations (9.9) and (9.10), the mean optical depth in the ionising photons from the inner to the outer edge of the ionised region is typically 6-8, so the optical depth at the Lyman limit, $\tau_{\nu_0} \sim 30$. This is sufficiently high that the diffuse photons will be re-absorbed within the nebula close to where they were produced.

We therefore introduce the idea of the *on-the-spot approximation*, which avoids the tedious business of tracing the details of the radiative transfer of the diffuse photons across the nebula. In this approximation, the *Baker & Menzel's Case B*, the recombinations to the ground state are ineffective in reducing the state of ionisation, since they produce a diffuse photon which promptly ionises another hydrogen atom in the vicinity. In Case B, therefore, the state of ionisation of the plasma is somewhat higher, and we compute the effective recombination coefficient as:

$$\alpha_{eff}^{H} = \alpha_{tot}^{H} - \alpha_{1S}^{H} \tag{9.15}$$

This reduces the recombination rate to about 60% of its *Baker & Menzel's Case A* value, the case where the nebula is so optically thin that all the photons produced by direct recombination to the ^1S level are free to escape the nebula.

The fact that normal H II regions are in Case B, and are therefore optically thick in the Lyman continuum photons ensures that they are also optically thick in all the Lyman series photons as well, and especially, in Ly α. This has an observable effect on the ratios of the Balmer lines; the *Balmer Decrement*.

9.1.3 Ionisation Fronts

H II regions are not born instantaneously. It takes time for the newly-formed exciting star to switch on its EUV radiation field and for the ionised region to expand towards its Strömgren radius. During this period, possibly only a small fraction of the EUV photons emitted by the central source are being used to maintain the ionisation in the nebula. In this circumstance, much of this radiation reaches the boundary of the ionised region unabsorbed, and these photons are available to push an *ionisation front* through the neutral medium, at a velocity given by equation (9.6); $\frac{dx}{dt} = q$.

During the initial evolution of the ionisation front, the rate of advance of the ionised region is equal to the ionisation parameter q at the boundary, which is determined by the density, the radius, the number of photons produced by the central star, and the number of these that are being used to maintain the ionisation in the body of the H II region. Let us assume that the ionised part of nebula is fully ionised, and that it is also in ionisation equilibrium, so that the number of ionisations equals the number of recombinations, locally. In this case:

$$4\pi r^2 n \frac{dr}{dt} = S_* - \frac{4\pi}{3}\alpha\left(T_e\right)n^2 r^3 \tag{9.16}$$

First, let us make this equation dimensionless by substituting the dimensionless distance variable $\xi = r/R_s$ for r, and the dimensionless time variable $\tau = t/\tau_s$, where R_s is the Strömgren radius given in equation (9.12) and τ_s is a *"Strömgren timescale"* defined as $\tau_s = 1/\alpha\left(T_e\right)n$. This time is, in fact,

the recombination timescale for the ionised plasma, $\tau_s \equiv \tau_{rec} \sim 10^5/n$ years. With these substitutions quation (9.16) then simplifies to:

$$\frac{d\xi}{d\tau} = \frac{(1 - \xi^3)}{3\xi^2} \tag{9.17}$$

which has the solution:

$$\xi = (1 - \exp[-\tau])^{1/3} \tag{9.18}$$

This equation shows (as it should) that initally, the radius of the ionised region is zero. However, equation (9.17) implies that the initial advance of the ionisation front is very rapid, and its speed varies as the inverse square of the radius. In reality, of course, the ionisation front cannot move faster than the speed of light. Indeed, until τ is greater than unity, even the basic assumption of ionisation equilibrium made in equation (9.16) is invalid. By this time the H II region has already grown to 86% of its Strömgren radius. At this time, the characteristic velocity of the ionisation front is, from equation (9.17), $v_{IF} = 0.09 R_s/\tau_{rec}$. The actual velocity depends on the nature of the exciting source, and on the density. If the central star is an O star, then typically, $S_* \sim 10^{49}$ sec^{-1}. With $S_{49} = S_*/10^{49}$ sec^{-1} we have:

$$v_{IF(\tau=1)} \sim 57 \, (S_{49}/n)^{1/3} \, \text{km s}^{-1} \tag{9.19}$$

The process of ionisation and heating of the plasma results in a large jump in the gas pressure. Since the temperature of the plasma goes from around 10^2 K to around 10^4 K during the ionisation, the gas pressure is of order a hundred times larger in the ionised plasma than in the atomic material ahead of the ionisation front. However, an ionisation front velocity of the size given by (9.19) is very much larger than the sound speed in either the ionised or the atomic gas ahead of the ionisation front. Thus, neither the ionised nor the neutral gas can react dynamically to the increased pressure in any significant way.

In a fast ionisation front, provided that we can neglect the thermal pressure in the neutral material, the ram pressure of the gas entering the front is matched by the sum of the gas pressure and the ram pressure of the ionised gas leaving the front. Thus the final velocity of the gas relative to the ionisation front is lower than the initial velocity, and the gas is compressed in the passage through the front. Such an ionisation front is termed *R-type* (meaning that the gas is rarefied ahead of the front). However, for a very fast ionisation front where the ionisation front velocity is very much larger than the sound speed in either the ionised or the atomic gas ahead of the ionisation front, neither the ionised nor the neutral gas can react dynamically to the increased pressure in any significant way. Such an ionisation front therefore produces only a small change in the density across it, and is termed a *weak R-type* front.

As the H II region expands towards its Strömgren radius, the velocity of the ionisation front continues to fall, until it approaches the sound speed in the ionised gas, but is still highly supersonic with respect to the atomic medium. In the example given above this happens after about two recombination times (or about $2 \times 10^5/n$ years, which is still very short compared with the lifetime of an O star; $\sim 5 \times 10^6$ years). Now the plasma attempts to adjust to the steep pressure gradient across the ionisation front, and the density change across it is much greater. At a certain velocity the hot plasma starts to push a strong compression shock into the atomic gas ahead of it. The exact velocity at which this happens will be derived in Section (9.1.4). Such an ionisation front is called an *D-critical* front.

As the expansion slows still further, the ionised region becomes detached from a strong shock which is propagating into the neutral gas and we have a *D-type* (dense) front, in which the gas ahead of the ionisation front is now denser than the ionised gas behind it. Since the expansion velocity is decreasing, the strength of this shock decays over time. Cooling may produce a shock which is approximately isothermal, and so the neutral gas is compressed by a factor of order one hundred. The pressure in the post-shock gas is now more-or-less matched to the pressure in the ionised region. By this time, the expansion velocity is subsonic with respect to the ionised plasma, so that the pressure gradients in the ionised plasma are much reduced. Because of the large density change, this type of front is dubbed *strong D-type*.

Eventually, the expansion velocity falls to the point where the motions in both the ionised and neutral phases are subsonic. By this time dynamical adjustments have allowed the pressure in the ionised plasma $(2n_2kT_2)$ to fall towards the pressure in the neutral plasma (nkT_1) so that the final Strömgren radius, $\left(3S_*/4\pi\alpha\left(T_e\right)n_2^2\right)^{1/3}$, is very much greater than the initial Strömgren radius. However the time required to reach this final configuration is much longer than the lifetime of the ionising star in most cases. In any event, the dynamical effects generated by the energetic stellar winds of hot stars are much more important than the dynamical effects generated in ionisation fronts at these late stages.

9.1.4 Jump Conditions in Ionisation Fronts

Advanced Topic

In this section, we will consider the gas dynamics across the ionisation front in a little more detail. The theory and the nomenclature were first worked out by Kahn (1958).

The thickness of the ionisation front is determined by equation (9.1), which, as we have already seen, is only about 1% of the intial Strömgren radius. In steady flow and in the frame of reference of the ionisation front, we may therefore treat the ionisation front as a discontinuity in the flow parameters, and apply the appropriate jump conditions across it.

As equation (9.6) makes clear, the motion of an ionisation front is controlled by the flux of ionising photons into it (or, equivalently, the ionisation parameter immediately before it). The mass flux through the front is therefore determined by the photon flux, and is not an arbitrary parameter, as it is in the case of shock waves:

$$\rho_0 v_0 = \rho_1 v_1 = m_H S \tag{9.20}$$

The gas pressure may be quite different on either side of the ionisation front because of the large increase in the temperature of the gas as it becomes ionised. On the other hand, we must also take the radiation pressure acting across the ionisation front into account. The momentum equation is therefore:

$$P_0 + \rho_0 v_0^2 = P_1 + \rho_1 v_1^2 + P_{rad} \tag{9.21}$$

or more explicitly:

$$\rho_0 \left(\frac{c_0^2}{\gamma} + v_0^2 \right) = \rho_1 \left(\frac{c_1^2}{\gamma} + v_1^2 \right) + \frac{h \langle \nu \rangle S}{c} \tag{9.22}$$

where we have substituted for the sound speed $c^2 = \gamma P/\rho$. For the radiation pressure term to be dominant, the radiation pressure must be comparable to the gas pressure in the ionised plasma, which condition gives $q \gtrsim \lambda m_H c_1^2/h$. Thus, the ionisation parameter q has to be greater than about 10^9 ($\mathcal{U} = 0.03$). This condition is not usually satisfied in normal H II regions of planetary nebulae, but radiation pressure may become important in photoionised regions near active galactic nuclei, or in the shells of gas ejected in nova explosions.

Ignoring the production of radiation within the ionisation front itself, but allowing for the heat energy produced by each photoionisation, the of energy conservation can be written:

$$\frac{\gamma}{\gamma - 1} \frac{P_0}{\rho_0} + \frac{v_0^2}{2} = \frac{\gamma}{\gamma - 1} \frac{P_1}{\rho_1} + \frac{v_1^2}{2} - \frac{h \langle \nu - \nu_0 \rangle S}{m_H} \tag{9.23}$$

or, substituting for the sound speed and re-arranging:

$$c_1^2 = c_0^2 + \frac{\gamma - 1}{2} v_0^2 - \frac{\gamma - 1}{2} v_1^2 - (\gamma - 1) \epsilon \tag{9.24}$$

where ϵ is the mean kinetic energy per unit mass liberated in the photoionisation process.

Equations (9.20) (9.22) and (9.24) provide the general solution for the motion of gas across a planar ionisation front. For simplicity, let us neglect the radiation pressure term. Then, eliminating the density dependence between the first two equations, we have:

$$\left(\frac{v_1}{v_0} \right)^2 - \left[1 + \frac{1}{\gamma \mathcal{M}_0^2} \right] \left(\frac{v_1}{v_0} \right) + \frac{1}{\gamma} \frac{1}{\mathcal{M}_1^2} = 0 \tag{9.25}$$

with the Mach number of the ionisation front in the undisturbed and ionised material being, respectively:

$$\mathcal{M}_0 = v_0/c_0 : \mathcal{M}_1 = v_0/c_1 \qquad (9.26)$$

$\mathcal{M}_0$ is defined by the initial conditions, and $\mathcal{M}_1$ can be determined from the energy equation (9.24). The solution for v_1/v_0 given by equation (9.25) must be both real and positive, which restricts the range of Mach numbers over which solutions are possible. Recognising that, thanks to the large increase in the temperature across the ionisation front $\mathcal{M}_0 \gg \mathcal{M}_1$, then the existence of a real solution for large Mach numbers (R-type solution) requires that the ionisation front must have, to first order:

$$\mathcal{M}_1 > \frac{2}{\gamma^{1/2}} : \text{or equivalently } v_R > \frac{2c_1}{\gamma^{1/2}} \qquad (9.27)$$

and the existence of a positive solution for small Mach numbers (D-type solution) requires that, to first order:

$$\mathcal{M}_1 > 2\gamma^{1/2}\mathcal{M}_0^2 : \text{or equivalently } v_D < \frac{c_0^2}{2\gamma^{1/2}c_1} \qquad (9.28)$$

Thus, since typically, $c_1 \sim 12$ km s^{-1}, the evolving ionisation front slows from weak-R to its *R-critical* value, $v_{R-crit} \sim 16$ km s^{-1} at which point it detaches a strong (approximately) isothermal shock into the neutral medium, which compresses the gas sufficiently to slow the ionisation front to its *D-critical* value. By the time that this transition occurs, the H II region has already expanded to about 95% of its initial Strömgren radius. The motion of shock is governed by the internal pressure in the H II region. Since the shock velocity, $v_s \sim c_1 \sim 12$ km s^{-1} and $c_0 \sim 1$ km s^{-1} the shock can be regarded as strong. Additionally, we can take the post-shock cooling to be strong enough that the shock can also be regarded as isothermal. Therefore:

$$P_{ram} = \rho_0 v_s^2 = m_H n_0 \left(\frac{dr}{dt}\right)^2 \text{ and}$$
$$P_{ion} = c_1^2 \rho_1 = m_H n_1 c_1^2 = P_{ram} \qquad (9.29)$$

However, because the H II region is already very close to its Strömgren radius at the time of shock detachment, the number of recombinations in the nebula ionisations must be approximately the same as the number of ionising photons produced by the central star:

$$S_* = \frac{4\pi}{3}\alpha(T_e)n_1^2 r^3 \qquad (9.30)$$

Equations (9.29) and (9.30) give the equation of motion of the shock front (behind which the ionisation front follows very closely):

$$r^{3/2} \left(\frac{dr}{dt} \right)^2 = c_1^2 R_s^{3/2} \tag{9.31}$$

where we have made use of the definition of the Strömgren radius given by (9.12). Substituting for the dimensionless distance variable of equation (9.17); $\xi = r/R_s$, and defining a new dimensionless time variable based on the sound-crossing timescale of the initial Strömgren sphere, $\tau_s = c_1 t/R_s$, equation (9.31) simplifies to:

$$\xi^{3/4} \frac{d\xi}{d\tau_s} = 1 \tag{9.32}$$

As pointed out above, the boundary condition $\xi = 1$; $\tau_s = 0$ is valid to a good level of approximation so that this equation can be integrated to give:

$$
\begin{aligned}
\xi &= (1 + 7\tau_s/4)^{4/7} \\
\frac{d\xi}{d\tau_s} &= (1 + 7\tau_s/4)^{-3/7}
\end{aligned}
\tag{9.33}
$$

Since the sound-crossing timescale of a typical H II region ($r \sim 10$ pc) is of order 10^6 years, this equation shows that the time taken to approach the final Strömgren radius (at which the internal pressure of the H II region matches that of the ISM) is many times the lifetime of the central star. The star will have blown up as a Type II supernova long before then.

9.1.5 Photoionisation of Neutral Condensations

Advanced Topic

H II regions are not homogeneous. They all contain lumps, condensations, partially ionised globules (PIGs!), and filaments of dense atomic and molecular gas left behind the main ionisation front. On the large scale, these may result from inhomogeneities in the placental cloud of neutral and atomic hydrogen or through crinkling instabilities in the ionisation front. On the small scale, these are associated with the parent clouds and the gaseous disks associated with the formation of low-mass main sequence stars and their planetary systems - stripped bare by the passage of the ionisation front when the exciting star of the H II region turns on. Very beautiful examples of both of these classes of condensation have been revealed by the HST.

Once formed, a neutral condensation of gas is surprisingly resilient to the destructive effects of photoionisation. This is because the layer of ionised gas streaming off the condensation serves to shield the cloud. The photoionising field at the ionisation front is weakened because much of it has been used in maintaining the ionisation of the flow region. A second factor aiding in the survival of such clouds is the fact that the recoil momentum of the ionised gas flowing from the ionisation front serves to compress the neutral gas ahead of it. This reduces the cross-section of the cloud to the ionising radiation field.

The theory of the photoionisation by a diffuse photon field of a neutral condensations in an H II region was worked out many years ago by Dyson (1968), and we follow his treatment here. The difference between a normal ionisation front and the ionisation front around a neutral condensation is that the photoionising field reaching the ionisation front is not a free variable, but instead is controlled by the radiative transfer through the ionised flow region. The radiative transfer through the flow region, the dynamics of the flow region and the structure of the ionisation front are therefore coupled, and have to be solved in a self-consistent way.

Dyson simplified the problem by making the eminently reasonable assumption that the ionised flow region is isothermal, is in photoionisation equilibrium, and has a steady-flow solution. The requirement for this to be satisfied is that the recombination timescale in the ionised flow is short compared with the dynamical timescale of the outflow. This is easily satisfied for the type of cloudlets seen in Orion. The dynamical timescale $\tau_{dyn} \sim 300 r_{16} \mathcal{M}^{-1}$ years, where $\mathcal{M}$ is the Mach number of the flow and r_{16} is the cloud radius in units of 10^{16} cm, while the recombination timescale $\tau_{rec} \sim 10^5/n$ years, as has been already shown. Therefore, the recombination timescale is less than the flow timescale for any flow with hydrogen number density $n > 300 r_{16} \mathcal{M}^{-1}$ cm^{-3}. To solve the steady-flow requires an inner boundary condition. This is obtained by making the assumption that in a self-shielded flow the ionisation front has to be *D-critical*. In this case, the gas streams away from the ionisation front at the sound speed in the ionised medium, c_1. This is a reasonable assumption, because such a flow satisfies the *Jouguet-point* condition (see Kaplan, 1966). In the D-critical case, the following relationships are satisfied:

$$v_0 = \frac{c_0^2}{2c_1} : v_1 = c_1 : \frac{\rho_1}{\rho_0} = \frac{c_0^2}{2c_1^2} \qquad (9.34)$$

After ionisation, the input of energy from the photon field which keeps the temperature constant also allows the gas to perform the PdV work of expansion, and to continually accelerate in the flow region. The steady-flow conditions of mass and momentum conservation in an isothermal outflow provide the velocity and density profile of the outflowing gas:

$$\frac{r}{r_c} = \mathcal{M}^{-1/2} \exp\left[\frac{\mathcal{M}^2 - 1}{4}\right]$$
$$\frac{\rho}{\rho_1} = \exp\left[-\frac{\mathcal{M}^2 - 1}{2}\right] \qquad (9.35)$$

where r_c is the radius of the neutral condensation (or radius at the ionisation front) and $\mathcal{M}$ is the Mach number of the flow.

The radiation field, initially isotropic at infinite distance, $S_\infty/2\pi$ photons cm^{-2} sr^{-1}, is attenuated through the flow as a function of both optical depth and of angle. Most of that attenuation occurs close to the ionisation

front. For example, by the time $r \sim 2r_c$, equation (9.35) shows that the density is only a tenth of that at the ionisation front, and the recombination rate is reduced by a factor of a hundred. We may therefore adopt the plane-parallel slab approximation to solve for the local radiation field at radius r; $S(r)$:

$$S(r) = S_\infty \int_{\theta=0}^{\pi/2} \exp[-\tau(r)\sec\theta]\sin\theta d\theta \qquad (9.36)$$

where $\tau(r)$ is the local optical depth in the ionising continuum. Close to the ionisation front, the optical depth is large, so that significant contributions to the local ionising field come only from small θ. With this additional approximation,

$$S(r) = S_\infty \frac{\exp[-\tau(r)]}{\tau(r)} \qquad (9.37)$$

Since the number of new photoionisations has to match the number of recombinations $(\alpha(T_e)n(r)^2)$ at each radial shell in the flow then:

$$S(r)\frac{d\tau(r)}{dr} = -\alpha(T_e)n(r)^2 \qquad (9.38)$$

using equation (9.37), integrating and substituting numerical values, this gives the the (self-consistent) density of the ionised flow at the ionisation front:

$$n(r_c) = 1.94\text{x}10^6 S_\infty^{1/2} r_c^{-1/2} \text{cm}^{-3} \qquad (9.39)$$

For the set of parameters which characterises the environment and the condensations in the Orion nebula (number of ionising photons produced by the exciting star, $S_* \sim 10^{49}$ s^{-1}, $S \sim 10^{12}$ cm^{-2} s^{-1}, mean density of the substrate nebula $\langle n \rangle \sim 10^3$ cm^{-3}, and $r_c \sim 10^{16}$ cm) equation (9.39) implies that $n(r_c) \sim 2 \times 10^4$ cm^{-3}, which from equation(9.35) suggests the typical densities in the neutral inclusions are $\sim 10^7$ cm^{-3}.

The outflow region is terminated when it interacts with the ionised substrate of the nebula. This occurs when the ram pressure in the flow matches the thermal pressure in the surrounding medium. At this point in the flow, an (isothermal) termination shock is formed, and the under-dense outflow is compressed by a factor of $\mathcal{M}^2$ to match the density of the substrate and to be brought to rest with respect to it. From equation (9.35), therefore:

$$\frac{\langle n \rangle}{n(r_c)} = \mathcal{M}^2 \exp\left[-\frac{\mathcal{M}^2-1}{2}\right] \qquad (9.40)$$

which with the parameters we have taken for Orion gives $\mathcal{M} \sim 3.5$, $r \sim 10r_c \sim 10^{17}$ cm. In practice, the substrate will not usually be stationary with respect to the flow, so a low Mach number bow-shock structure is formed in the substrate at the termination of the outflow. This will induce flows and

turbulence with velocities of order of the sound speed. Both the spatial scale ($\sim 10''$ at the distance of Orion, ~ 500 pc) and the magnitude expected of this turbulence ($\sim 10 - 15$ km s^{-1}) is in good agreement with the high-resolution data of Wilson *et al.* (1959). It is clear that ionised flows from dense inclusions may drive both the large temperature and density fluctuations which apparently characterise H II regions as a whole.

9.2 H II Regions with Heavy Elements

A real H II region will always contain heavy elements. The reason we have been able to go so far by considering only the hydrogen within the H II region is because atomic hydrogen is the dominant source of opacity for frequencies at, and somewhat above, the Lyman limit, ν_0. However, helium may also contribute significantly to the opacity for photon energies high enough to ionise this element; $\nu \geq 1.8\nu_0$. Heavy elements provide some contribution as well, although this is much smaller than either hydrogen and helium, except for frequencies approaching and into the X-ray region, where K and L -shell ionisation of heavy elements provides the dominant opacity source.

Of course, our treatment so far has skipped lightly around the question of the equilibrium temperature, which, according to equation (9.2) is dominated by the collisional excitation cooling in emission lines of the heavy elements. Proper photoionisation modelling of H II regions would solve the full set of coupled ionisation - cooling equations, including the radiative transfer of both the direct and diffuse radiation fields and taking proper account of the geometrical dilution and other physical and geometrical effects such as clumping of the ionised matter and its distribution with respect to the central source. This is done correctly in such modelling codes as CLOUDY (Ferland,1993), MAPPINGS III (Sutherland & Dopita, 1993) and in a number of other codes as well. However, in the spirit of our philosophy of seeking physical insight through simplification, let us continue (for the time being) to assume that the electron temperature is about 10^4K and consider the ionisation structure that results in a nebula composed of only hydrogen and helium.

9.2.1 Ionisation Structure of Hydrogen & Helium Nebulae.

If hydrogen and helium are the only sources of opacity, then at any point in the ionised volume, the element of optical depth is simply:

$$d\tau_\nu = -\kappa_\nu dx = (n_{H^0}\sigma_H(\nu) + n_{He^0}\sigma_{He}(\nu) + n_{He^+}\sigma_{He^+}(\nu))\,dx \qquad (9.41)$$

it is obvious, therefore, that the radiation field can be divided into three parts, or bands according to how many of these opacity sources are active at any particular frequency:
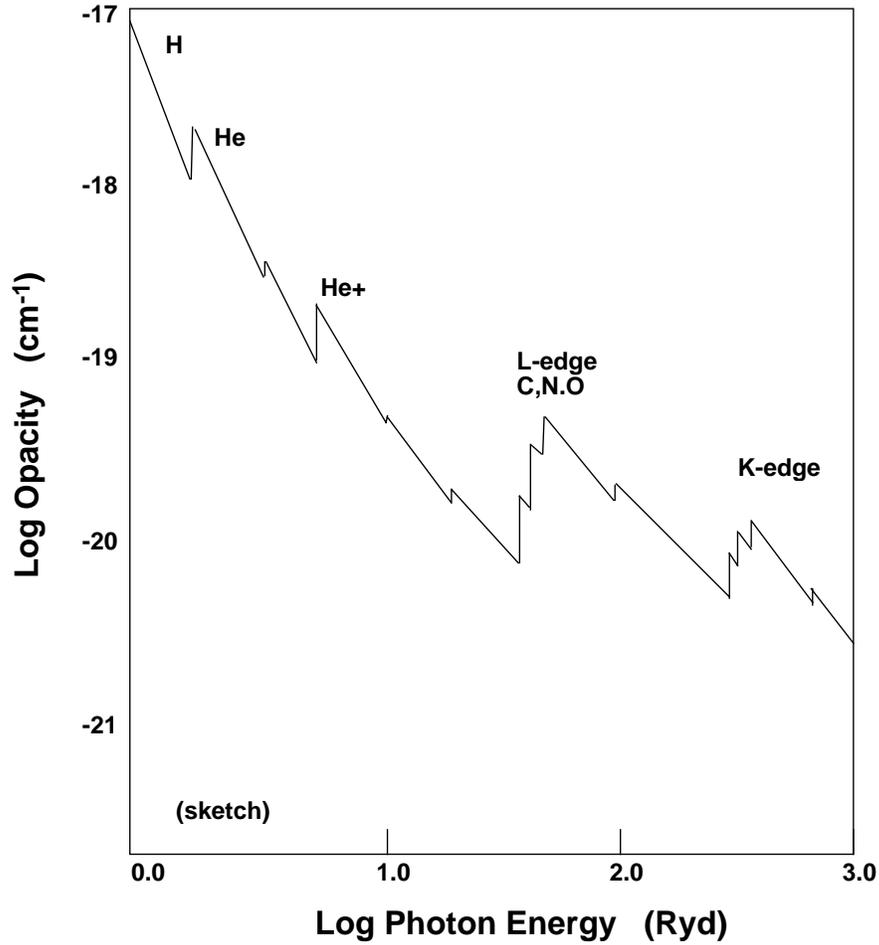
Fig. 9.2. The opacity of (un-ionised) atomic interstellar gas of solar composition as a function of photon frequency in Rydberg energy units (13.6 eV or 3.266×10^{15} Hz).

$$\text{A-band} \; : 1.0\nu_0 \leq \nu < 1.8\nu_0 : \text{H}^0 \text{ -ionising}$$
$$\text{B-band} \; : 1.8\nu_0 \leq \nu < 4.0\nu_0 : \text{H}^0, \text{He}^0 \text{ -ionising}$$
$$\text{C-band} \; : 4.0\nu_0 \leq \nu : \text{H}^0, \text{He}^+ \text{ -ionising}$$

where ν_0 is the threshold for ionisation of hydrogen.

In normal H II regions around young, hot stars, the number of C-band photons are negligible, since the stellar effective temperatures are so low that all such photons are absorbed in the passage through the stellar atmosphere. However, for the radiative transfer in the nebula, the absorption of B-band photons by helium cannot be neglected. Although the typical abundance

of helium is a tenth by number with respect to hydrogen, this abundance difference is made up by the much greater photoionisation cross-section of helium to the B-band photons. At the threshold of ionisation of helium this cross-section is almost ten times as great as that of hydrogen at the same frequency; $\sigma_{He}(1.8\nu_0) \sim 9.9\sigma_H(1.8\nu_0)$, and the ratio between the two increases with frequency.

In the hydrogen nebula, the diffuse radiation field was fairly unimportant, since the only source of diffuse ionising photons was recombination to the ground state, which can be accounted quite well by the on-the-spot approximation. Helium likewise produces a helium-ionising radiation field by direct recombinations to the $1\ ^1S$ ground state of He I, but this is less important than in hydrogen, because only 1/4 of all recombinations occur to singlet levels, the remaining 3/4 occurring to levels in the triplet series. For this transition, use of the on-the-spot treatment is a rather poor approximation to the radiative transfer problem, since the nebular opacity in the singlet states of helium is much lower (by a factor of about 40) than for hydrogen.

Helium not only produces ionising photons by direct recombination to the ground state, but also in transitions between its $n = 2$ and $n = 1$ states. All of the recombinations to the triplet terms will produce a band A photon in the $2\ ^3P$ - $1\ ^1S$ transition. Of the remaining recombinations to excited states of the singlet series, about 2/3 will produce a hydrogen-ionising photon in the $2\ ^1P$ - $1\ ^1S$ transition, and about 56% of the recombinations to the $2\ ^1S$ state will produce a hydrogen-ionising photon in the $2q$ continuum between the $2\ ^1S$ and the $1\ ^1S$ terms. Summing all these contributions, we find that (effectively) every recombination of helium produces an A-band photon.

In this circumstance, to a reasonable approximation, the helium radiative transfer problem can be decoupled from the hydrogen transfer problem, and we can therefore talk about a helium Strömgren sphere the size of which is controlled only by the radiative transfer of the band B photons. Thus, if the star produces $S_* = S_A + S_B$ photons, if helium is present at an abundance $Z_{He} = N(He)/N(H)$, and the number density of hydrogen atoms is n_H, then the helium ionisation balance equations within the Helium Strömgren sphere are written (*c.f.* equation (9.13));

$$\frac{(\chi_H + Z_{He}\chi_{He})\chi_{He}}{1 - \chi_{He}} = \frac{\langle\sigma_{He}\rangle}{\alpha_{He}(T_e)} \frac{S_B \exp\left[-\tau_B(r)\right]}{4\pi r^2 n_H} \tag{9.42}$$

where:

$$\tau_B(r) = \int_0^r n\left[\langle\sigma_H\rangle(1 - \chi_H) + Z_{He}\langle\sigma_{He}\rangle(1 - \chi_{He})\right] dr \tag{9.43}$$

$\langle\sigma_H\rangle$ being the frequency-weighted mean photoionisation cross-section of hydrogen in the B-band photons, and $\langle\sigma_{He}\rangle$ being the corresponding value for helium. The (approximate) hydrogen ionisation balance equations are:

$$\frac{(\chi_H + Z_{He}\chi_{He})\chi_H}{1 - \chi_H} = \frac{\langle \sigma \rangle}{\alpha_H (T_e)} \frac{S_* \exp\left[-\tau_A (r)\right]}{4\pi r^2 n_H} \tag{9.44}$$

where $\langle \sigma \rangle$ is the frequency-weighted mean photoionisation cross-section of hydrogen in the A-band photons, including the He diffuse field photons, and:

$$\tau_A (r) = \int_0^r n \langle \sigma \rangle (1 - \chi_H) dr \tag{9.45}$$

These last two equations implicitly assume that the hydrogen-ionising diffuse field produced by helium is absorbed by hydrogen according to the on-the-spot approximation, which may not be particularly exact.

Equation (9.43) shows that helium will absorb most of the B-band photons when:

$$Z_{He} \frac{\langle \sigma_{He} \rangle}{\langle \sigma_H \rangle} \frac{(1 - \chi_{He})}{(1 - \chi_H)} > 1 \tag{9.46}$$

since $Z_{He} \langle \sigma_{He} \rangle / \langle \sigma_H \rangle \sim 1$ in a plasma with the solar abundances, this is condition tantamount to the requirement that $\chi_{He} < \chi_H$.

The helium Strömgren sphere will be smaller than the Hydrogen Strömgren sphere, if, at the inner edge of the nebula the condition $\chi_{He} < \chi_H$ is also satisfied. This is obvious from the form of the solution of the ionisation balance equation *see*, for example Figure(9.01); the lower the initial ionisation state, the smaller the effective ionisation parameter at the inner boundary, and the smaller (or thinner) the extent of the ionised region. Dividing equation (9.44) by (9.42) with the approximation that the fractional ionisation is of order unity then yields the condition on the ratio of the B-band photons to the total number of ionising photons:

$$\frac{S_B}{S_*} = \frac{S_B}{S_{A+B}} < \frac{\alpha_{He} (T_e) \langle \sigma \rangle}{\alpha_H (T_e) \langle \sigma_{He} \rangle} \tag{9.47}$$

since this ratio is a strong function of the effective temperature of the ionising source, this shows that the helium Strömgren sphere grows to fill the hydrogen Strömgren sphere as the temperature is increased. Detailed calculations show that they become effectively the same for stellar temperatures above about 40,000 K; *see* Figure (9.3).

Clearly, the helium Strömgren sphere cannot grow larger than the hydrogen Strömgren sphere, since helium is ionised wherever there are B-band photons. At high stellar effective temperatures, most of the helium ionising photons are absorbed by hydrogen.

From a practical viewpoint, it is very important to know the ratio of the hydrogen and helium Strömgren spheres, since the measurement of the primordial helium abundance, which is a critical cosmological parameter, depends very critically upon the measurement of the ratios of helium and hydrogen recombination lines in low-abundance H II regions. Line ratios such
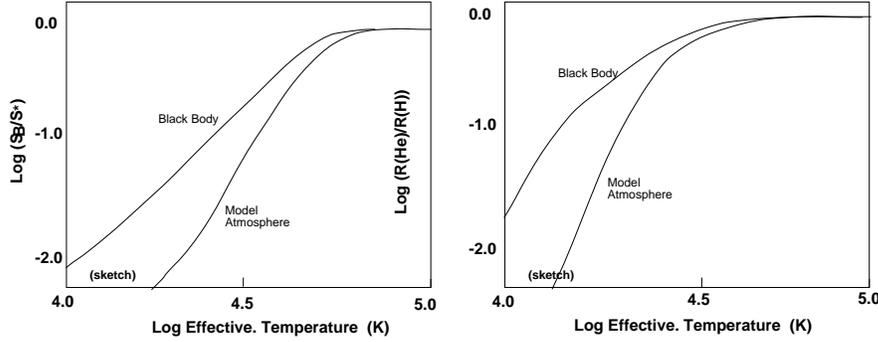
Fig. 9.3. The ratio of helium-ionising photons to hydrogen-ionising photons as a function of temperature (left) and the resulting ratio of the Strömgren radii (right), from detailed photoionisation modelling with MAPPINGS III.

as He I $\lambda4471$Å/Hβ orHe I $\lambda5876$Å/Hα are used for this purpose. We therefore effectively measure $N(He^+)/N(H^+)$ when we would rather determine $N(He^+ + He^0)/N(H^+)$ in the ionised region. The correction for the neutral helium is therefore rather important. To do this properly requires not only a detailed stellar atmosphere model, which tells us the intrinsic flux ratio of the A-band to B-band photons, but also a detailed nebula model, which takes proper account of the temperature and the ionisation structure of the nebula.

When band-C photons are present, for stellar effective temperatures above about 80000 K, then a He^{++} Strömgren sphere is also formed, nestled inside the He^+ sphere. The presence of such a sphere is signalled by the appearance of recombination lines such as He II $\lambda4686$Å or He II $\lambda1640$Å. As the temperature increases, this zone also expands and tends to fill the hydrogen Strömgren sphere. The limiting He II $\lambda4686$Å/Hβ flux ratio (and indeed, any similar ratio) is set by the ratio of the abundances, and the ratio of the effective recombination coefficients to produce the lines. In practice, the limiting He II $\lambda4686$Å/Hβ ratio lies in the range $\sim 0.5 - 0.7$, depending on the helium abundance.

9.2.2 Nebular Structure with Heavy Elements

As mentioned above, it is the hydrogen and helium which dominate the opacity, and consequently which control the local ionising field in the nebula. Since the ionisation state of the heavy elements is determined by the local ionising radiation field, then this must also be closely correlated to the local ionisation state of hydrogen and helium. As a consequence, we can expect that those elements which have ionisation potentials above 4 Rydbergs are located in the He^{++} zone, those with ionisation potentials between 1.8 and 4 Rydbergs are to be found in the He^+ zone, those with ionisation potentials between 1 and 1.8 Rydbergs occur in the H^+/He^0 zone, and finally neutral species

and ions with ionisation potentials below 1.0 Rydbergs are co-extensive with the atomic hydrogen. This predicts that the dominant ionisation zones of the nebula for the most abundant elements and important coolants are as follows:

H I, He I   : C II, N I, O I, Ne I, S II
H II, He I   : C II, (C III), N II, O II, Ne II, S II, (S III)
H II, He II : C III, (C IV), N III, O III, Ne III, S III, (S IV, S V)
H II, He III : C IV, N IV, O IV, Ne III, S V, and higher
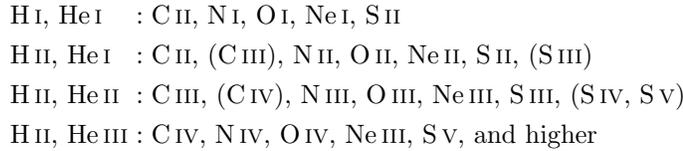
where non-dominant ionisation stages are indicated in parentheses.

Within each of these zones, but especially in the first three, the ionisation structure is appreciably modified by the charge-exchange reactions:

$$A^{(i+1)+} + H^0 \;\rightleftharpoons\; A^{i+} + H^+ + \Delta E \text{ and}$$
$$A^{(i+1)+} + He^0 \rightleftharpoons A^{i+} + He^+ + \Delta E$$
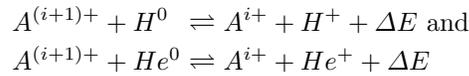
these, for instance, lock the ionisation ratio O II/O I so that it varies very much as H II/H I ratio, and serve to lower the O III/O II ratio in the intermediate ionisation zones.

The distinction between the atomic, A-band ionised and B-band ionised regions can be clearly seen in Figure (9.4) which is a full MAPPINGS III photoionisation model for an isochoric (constant density) region excited by a star with an effective temperature of 3.5×10^4 K.

9.2.3 Nebular Equilibrium Temperature

In an H II region, the local heat input is determined by the local rate of photoionisations of hydrogen and helium and the mean energy of the liberated photoelectrons, which is a function of the intensity and shape of the local ionising radiation field: $\epsilon(\mathcal{U}, T_{eff}, \tau)$. At low densities, the local energy loss rate is given by the cooling function appropriate to the local electron temperature, ionisation state and the heavy element abundance set; $\Lambda(T_e, Z)$. If, locally, we also have photoionisation equilibrium, then the photoionisation rate is equal to the recombination rate. Ignoring the small recombination heating term, the heat balance equation (given in equation (9.2) for the case where hydrogen supplies the photoelectric heating), can be then written:

$$\dot{Q} = \epsilon(\mathcal{U}, T_{eff}, \tau)\left(\alpha_H(T_e)n_{H^+}n_e + \alpha_{He}(T_e)n_{He^+}n_e + \alpha_{He^+}(T_e)n_{He^{++}}n_e\right)$$
$$= \Lambda(T_e, Z)nn_e \tag{9.48}$$

For plasmas with heavy element content below or about equal to solar, we can approximate $n \sim n_H + n_{He} = (1 + Z(He))n_H$, so that in the zone where only hydrogen is ionised (9.48) simplifies to:
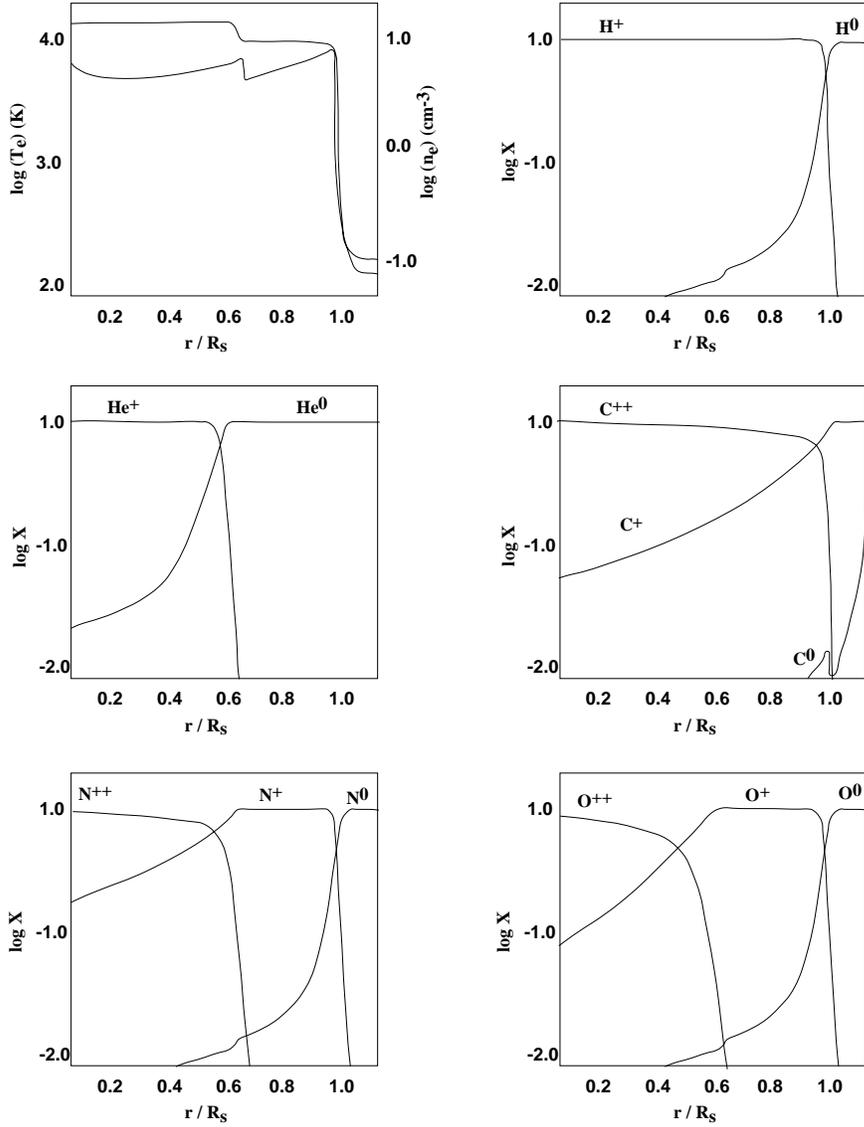
Fig. 9.4. The temperature, density, and ionisation structure of a model H II region illuminated by a star with an effective temperature of 35,000K. Note how the ionisation structure in the heavy elements echoes that of hydrogen and helium.
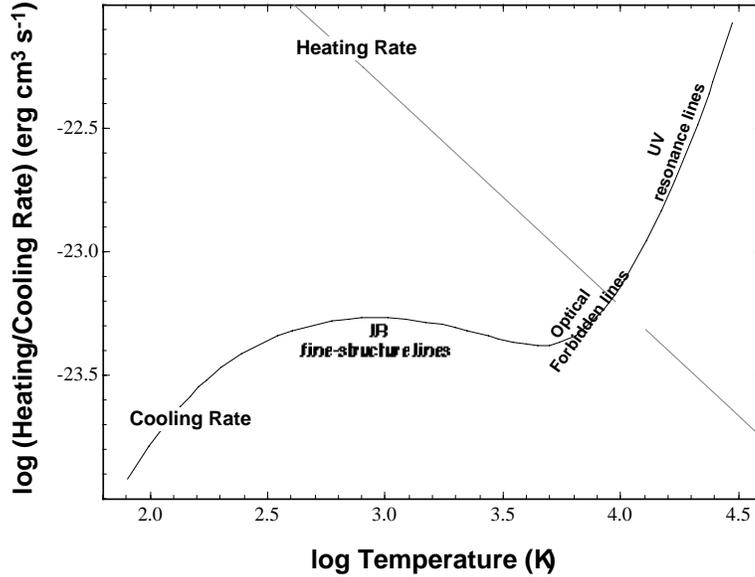
Fig. 9.5. The cooling function for a fixed ionisation produced by an O-star with $T_{eff} = 40000$K and with $q = 10^8$ cm s-1 is shown here as a function of electron temperature. The heating rate is related to the recombination rate, as described in the text. The equilibrium temperature is defined by the point at which these cross. Quite wide variations in the heating rate would produce only small changes in T_e, which remains in the vicinity of 10^4 K.

$$\epsilon(\mathcal{U}, T_{eff}, \tau)\alpha_H(T_e) = (1 + Z(He))\, \Lambda(T_e, Z) \qquad (9.49)$$

and within the zone where hydrogen is ionised and helium is singly ionised:

$$\epsilon(\mathcal{U}, T_{eff}, \tau)\,(\alpha_H(T_e) + Z(He)\alpha_{He}(T_e)) = (1 + Z(He))\, \Lambda(T_e, Z) \qquad (9.50)$$

Since the recombination rates are of order 2×10^{-13} cm^3 s^{-1} and the mean energy per photoionisation is about a Rydberg (2×10^{-11} ergs), the heating rate is typically about 4×10^{-24} erg cm^3 s^{-1}.

In Figure (9.5), we show the cooling function for fixed ionisation, the equilibrium ionisation in a O-star radiation field of 40000 K and with $q = 10^8$ cm s^{-1}. The ground-term fine structure splitting of the abundant heavy ionic species present produces many mid- and far- IR transitions with excitation energies of order 100 K, which accounts for the rapid rise in the cooling function. At temperatures of a few thousand the optical forbidden lines between the different terms with the same principal quantum number are excited, and the cooling function rises rapidly with temperature. Above 15000 K, approximately, the UV resonance lines come to dominate the cooling function, which continues to rise sharply with increasing temperature.

Figure (9.5) shows how this rapid increase in cooling rate with temperature, combined with the decreasing heating rate, serves to "thermostat" the H II region within a rather narrow range of electron temperatures over quite widely different heating rates, provided the temperature is above, or in the vicinity of 10^4 K. This is the case in all H II regions excited by O- or B-stars, provided the abundance of the heavy elements is solar or less. Such H II regions produce a rich forbidden-line spectrum in which the strengths of the forbidden lines exceeds the sum of the Balmer recombination lines of hydrogen at optical wavelengths.

When the heavy elements are present only at very low abundance, the electron temperature is so high that the cooling is dominated by the UV resonance and intercombination lines. In the optical, the forbidden lines approach their peak emissivity per atom (*c.f.* Figure (3.3)), but decreasing abundance decreases their intensity relative to the Balmer lines, and the recombination lines of hydrogen and helium come to dominate the spectrum. However, the [O III] lines remain an important coolant even at very low abundance. Eventually, when the abundance is lower than about a tenth to a thirtieth of solar, even these fade away.

For metal-rich OB stars and nebulae with metallicities in excess of about twice solar, the equilibrium electron temperature decreases rapidly with increasing cooling rate (or decreasing heating rate) in the temperature range $3.8 > \log T_e > 2.8$, approximately. Since metal-rich H II regions are too cool to excite the optical forbidden lines to any great extent, the optical spectrum is dominated by the recombination lines.

Since the forbidden lines reach a peak relative intensity at a particular abundance, any particular forbidden to Balmer line ratio is obtained for two values of the abundance. To identify which branch represents the correct solution, a measure of the electron temperature is required. This will be high for low chemical abundance, and low for high abundances.

Finally, note that the photoionisation equilibrium represented in Figure (9.5) is thermally stable, even where it is quite sensitive to changes in the heating rate. Instabilities will only occur where there is an unstable time-dependent feedback between the ionisation state and the cooling rate. This requires that there be an appreciable fraction of neutral hydrogen in the plasma.

9.3 Photoionisation by X-ray Sources

So far we have considered only the temperature and ionisation equilibrium of regions ionised by UV sources which have an energy spectrum $\nu F(\nu)$ which is effectively cut off at energies above about 100 eV. However sources much "harder" than this are observed. These include the very hottest planetary nuclei (PNn), which are observed as soft thermal X-ray sources, accreting white dwarf stars in binary systems, which are bright even in hard X-rays,

and various classes of active glactic nuclei (AGN). The theory of such regions was pioneered by Williams (1967) and MacAlpine (1972), and developed in the late 1970s and early 1980s by Koski (1978), Ferland & Netzer (1983) and Stasinska (1984), and continues to be an active field of research (*i.e.* Murayama & Taniguchi, 1998)

In H II regions ionised by a photon field which extends up to X-ray frequencies (even if it is not dominated energetically by the X-ray photons), the physics described in Section (5.3.2) comes into play. The opacity of the medium is now dominated by inner shell photoionisation processes in heavy elements. These, and inner shell photoionisation / Auger ionisation processes such as:

$$O^{2+}(1s^2 2s^2 2p^2 {}^3P) + h\nu \rightarrow O^{3+}(1s2s^2 2p^2 {}^2P) + e$$
$$O^{3+}(1s2s^2 2p^2 {}^2P) \qquad \rightarrow O^{4+}(1s^2 2s^2 {}^2P) + e$$
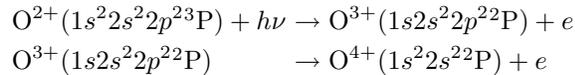
couple ionisation stages which differ by two (or more) electrons, and produce high-energy fast electrons capable of producing many collisional exciations or secondary ionisations in hydrogen or helium.

In the region of an X-ray ionised H II region which is closest to the source, these processes coupled with the high ionisation parameter produce a hot, highly ionised plasma with an intense and rich emission line spectrum at optical and UV wavelengths. In such a plasma the [Ne V]$\lambda\lambda$3364,3426Ålines are very strong, and the "coronal" lines of other species, so called because they were first seen in the corona of the Sun, are prominent. These include [Fe VII]$\lambda\lambda$5720,6087Å, [Fe X]λ6375Åand [Fe XI]λ7892Ålines, which require photon energies up to at least $h\nu \gtrsim 260$ eV. In some extreme AGN [Fe XIV] λ5303Åhas even been seen.

In the bulk of the H II region, however, the ionisation state is much lower, and ions such [O III] (which are found in normal H II regions), are the dominant ionisation stage. The difference between X-ray photoionised and normal H II regions is that the fast electrons generate collisional heating which produces a very rich optical and UV line spectrum. In general, the forbidden lines are systematically stronger in an X-ray photoionised nebula, and become stronger still as the spectrum of the central source becomes harder.

The most profound difference between an X-ray ionised region and a normal H II region is found near the outer boundary. In a normal H II region, the opacity of hydrogen and helium increases sharply near this boundary, producing the rapid transition between ionised and neutral plasma described in Section (9.1.1), above. In an X-ray ionised region, the photon field becomes very hard in the outer regions. Such a photon field is quite unaffected by the state of ionisation of hydrogen and helium, and the mean free path of the ionising photons is long. Although the local ionisation parameter is quite low, these residual photons are very energetic, and are capable of producing significant heating when absorbed. In addition, the energetic photoelectrons

produce many collisional excitations and secondary ionisations in atomic hydrogen.

The result of all of this is that the hard-EUV/ soft-X-ray photons produce an extended transition zone in which the temperature is maintained in the range 5000 to 10000 K, and in which hydrogen has a significant fractional ionisation (a few percent, typically). These conditions are ideal for producing strong emission lines in neutral species, or species with ionisation potentials below that of hydrogen, such as [O I]$\lambda\lambda$6300,6363Å, [N I]$\lambda\lambda$5198,5200Å, and [S II]$\lambda\lambda$6717,6731Å or $\lambda\lambda$4069,4076Å. Indeed, a high [O I] $\lambda\lambda$6300/Hα ratio, combined with the presence of lines of high excitation is an excellent diagnostic for the presence of a source of hard photons. In addition, the collisional excitations to the $n = 3$ level of hydrogen enhance the strength of Hα above its Case B recombination value, so that X-ray ionised regions are also characterised by a somewhat enhanced Balmer decrement.

9.4 Radio Continuum of H II Regions

The source of the radio continuum emissivity of H II regions is the thermal free-free continuum (sometimes called the *thermal bremsstrahlung*) described in Chapter 6. This continuum is produced by the acceleration of the free electrons during Coulomb scatterings caused by the heavy ions, and is limited at high frequencies by the requirement that the photon produced cannot exceed the kinetic energy of the electron being scattered. At radio wavelengths, the opacity (which may be large) is provided by the inverse process; free-free absorption.

When the nebula is very optically thick, the emergent specific intensity at any wavelength is simply the Planck Black-Body value appropriate to the plasma temperature; given by equation (4.12):

$$I(\nu) = B(\nu, T_e) = \frac{2h\nu^3}{c^2 \left(\exp\left[\frac{h\nu}{kT_e} \right] - 1 \right)}$$
$$\sim 2\nu^2 kT_e/c^2 : h\nu/kT_e \ll 1 \tag{9.51}$$

Here, the *Rayleigh-Jeans* approximation given in the second line is appropriate for H II regions (with electron temperatures of several thousand K) for radio frequencies of less than about 100GHz. In the limit of high optical depth, the equation of transfer through the nebula is very simple:

$$\frac{dI(\nu)}{ds} = -\kappa_\nu I(\nu) + j_\nu = 0 \tag{9.52}$$

These two equations gives the (*Kirchhoff's Law*) relationship between the emission and the absorption coefficient; $j_\nu = \kappa_\nu B(\nu, T_e)$. At the limit of high optical depth, the *brightness temperature*, T_B, defined as $I(\nu) = B(\nu, T_B)$

is identical to the electron temperature, and, from equation (9.51) the radio continuum spectrum is simply a power law in frequency with slope $+2$; $I(\nu) \propto \nu^2$.

The emission coefficient for the free-free interaction of electrons with ions of charge Z_i is:

$$j_\nu = \frac{8g_{ff}}{3} \left(\frac{2\pi}{3}\right)^{1/2} \frac{Z_i^2 e^6 \exp\left[-h\nu/kT_e\right]}{m_e^{3/2} c^3 \left(kT_e\right)^{1/2}} n_i n_e \tag{9.53}$$

$$= 5.44 \times 10^{-39} g_{ff} Z_i^2 n_i n_e T_e^{-1/2}$$

$$\text{x} \exp\left[-h\nu/kT_e\right] \text{erg cm}^{-3} \text{ s}^{-1} \text{ sr}^{-1} \text{ Hz}^{-1}$$

where at radio wavelengths, the free-free Gaunt factor can be approximated by:

$$g_{ff} = 11.96 T_e^{0.15} \nu^{-0.1} \tag{9.54}$$

Combining equations (9.51),(9.53) and (9.54) into the Kirchoff relationship then gives:

$$\kappa_\nu = 0.212 Z_i^2 n_i n_e \nu^{-2.1} T_e^{-1.35} \text{cm}^{-1} \tag{9.55}$$

from this equation it then follows that, for a fully-ionised plasma of solar composition, the optical depth for a column of length s for free-free absorption; $\tau_\nu = \kappa_\nu s$, is directly related to the emission measure of the nebula $\langle EM \rangle = \int n_e^2 ds$:

$$\tau_\nu \sim 0.08235 \nu^{-2.1} T_e^{-1.35} \langle EM \rangle \tag{9.56}$$

The interesting thing about this equation is that it shows that the nebula is optically thick at low frequencies, but becomes optically thin to free-free absorption at high frequencies.

For finite optical depth, and in the Rayleigh-Jeans limit of equation (9.51), the equation of transfer (9.52) is re-written in terms of the optical depth and the brightness temperature defined above:

$$\frac{dT_B}{d\tau_\nu} = -T_B + T_e \tag{9.57}$$

thus, the emergent specific intensity, expressed in terms of the brightness temperature is

$$T_B(\nu) = T_e\left(1 - \exp[-\tau_\nu]\right) + T_B(0) \tag{9.58}$$

where $T_B(0)$ is the brightness temperature of the sky background at this frequency (with specific intensity $I(0)$). In the optically thin (high frequency) limit, equation (9.56) gives (provided the Rayleigh-Jeans approximation is still valid);

$$I(\nu, T_e) - I(0) = 2\nu^2 \tau_\nu kT_e/c^2 \propto \nu^{-0.1} T_e^{-0.35} \langle EM \rangle \tag{9.59}$$

In summary: at low frequencies, the nebula is optically thick, the free-free radio continuum spectrum has a flux which increases as the square of the frequency, and from which the electron temperature can be determined. At high

frequencies, where $\tau_\nu < 1$, the flux varies as $\nu^{-0.1}$. At these high frequencies, the emission measure of the nebula can be obtained from a determination of the surface brightness in the radio continuum. This measurement provides an accurate result provided that the electron temperature has been already determined from the low-frequency spectrum, and provided that the radio telescope has got sufficient spatial resolution to properly resolve the nebula (including any small high-density inclusions which it might contain).

Note that the emission measure is closely related to the recombination rate in the nebula. Integrating the surface brightness over the whole nebula to give the *flux density*, $S(\nu)$, of the source, therefore gives a direct estimate of the ionisation rate of the central source (and hence either the mass of the nebula or the luminosity of the central source), provided that the distance can be determined (and with it, the ionised volume). Putting numerical values into equation (9.59):

$$S(\nu) = 27.18 \times 10^{-15} \nu^{-0.1} T_e^{-0.35} \left\langle n_e^2 \right\rangle \theta^3 D \text{ Jy} \qquad (9.60)$$

where the flux density is expressed in Janskys (10^{-26} W m^{-2} Hz^{-1}), and the frequency is in Hz, θ is the angular radius of the H II region in radians, D is is the nebular distance (cm).

Notes on Chapter 9

- The theory of the dynamics of H II regions by Dyson, J.E. & Williams, D.A. 1997 *The Physics of the Interstellar Medium, 2nd. Ed.*, Institute of Physics Publishing: Bristol, ISBN 0-7503-0460 (paperback) is an excellent place to start.
- The theory of ionisation fronts is carefully presented in Kaplan, S.A. 1966, *Interstellar Gas Dynamics, 2nd Revised Edition*, ed. F.D. Kahn, Pergamon Press: Oxford.
- A very complete discussion of the line and continuous spectra of H II regions is given in Aller, L.H. 1984, *Physics of Thermal Gaseous Nebulae*, Reidel:Dordrecht, ISBN 90-277-1814-8, and in Osterbrock, D. *The Astrophysics of Gaseous Nebulae*. A good exposition on radio observations of H II regions is by Brown, R.L. 1987, in *Spectroscopy of Astrophysical Plasmas* eds. A.Dalgarno & D. Layzer, CUP: Cambridge, ISBN 0-521-26315-8, p35. All these sources have been used in the preparation of this chapter.

Exercise 9.4.1. Near the edge of a (plane-parallel) nebula, the mean photoionization cross-section per hydrogen atom is σ, the mean energy given to the gas per photoionisation is Γ ergs, and hydrogen accounts for effectively all the photoionisations and the electrons which are present.

a. If the cooling rate of the plasma per unit volume is $n_e n_H \Lambda$ erg cm^{-3} s^{-1}, the effective recombination coefficient is α cm^3 s^{-1}, and the gas is in local photoionization equilibrium, give expressions for the fractional ionization of hydrogen $x = n_e/n_H$ and the equilibrium cooling rate as a function of the

dimensionless ionization parameter $\mathcal{U}$ defined in equation (9.7). **Hint:** Read section (9.1) to work out the ionization fraction. Pick the sensible solution, and then balance the heat gain against the heat loss.

b. (for advanced students) If the cooling rate of the plasma per unit volume is $n_e n_H \Lambda(T_e)$,erg cm^{-3} s^{-1}, where $\Lambda(T_e)$ can be approximated by the equation :

$$\Lambda(T_e) = 10^{-23} \exp\left[-\frac{1000K}{T_e}\right]\left(\frac{T_e}{10^4 K}\right)^{-1/2} + 10^{-21} \exp\left[-\frac{20,000K}{T_e}\right]\left(\frac{T_e}{10^4 K}\right)^{-1/2}$$

the effective recombination coefficient is $\alpha(T_e) = 10^{-13}\left(\frac{T_e}{10^4 K}\right)^{-1}$ cm^3 s^{-1}, the effective cross section per hydrogen atom is $\sigma = 10^{-18}$ cm^{-2}, and assuming local photoionization equilibrium, compute graphically using any method (FORTRAN, Pacal or C program, MATHEMATICA or MAPLE, pencil and paper *etc.*) the fractional ionization of hydrogen and the equilibrium temperate as a function of the ionization parameter $\mathcal{U}$ in the range $10^{-7} \leq \mathcal{U} \leq 10^{-4}$.

Exercise 9.4.2. An active galaxy nucleus has an ionizing luminosity of 10^{44} erg s^{-1}, and the mean photon energy is 50eV. It is surrounded by an extended photoionised region with a density profile $n(r) = 100[1 + (r/100\text{pc})^{-2}]$ cm^{-3}.

a. What is the size of the region that the active nucleus can ionize?

b. Assuming an inner radius of 30 pc for the ionized gas, is the ratio of the radiation pressure to the gas pressure at this radius, assuming that the gas temperature is 10^4 K?

c. Is this radiation pressure effective in accelerating the plasma at this radius outwards? Explain your reasoning.

Exercise 9.4.3. An $80M_\odot$ O star produces ionizing photons at a rate, S_*, of 10^{50} s^{-1}. The medium around the star consists of pure hydrogen, and has a constant density, $n_0 = 10$ cm^{-3}.

a. If the recombination coefficient in the ionized gas, $\alpha = 3 \times 10^{-13}$ cm^3 s^{-1}, what is the Strömgren radius, R_0 , of the HII region that it initially produces, assuming that the difference in pressure between the ionised region and the non-ionised region has not yet had time to produce any dynamical effects? **Answer:** 43.4 pc.

b. What would the final radius, R_1 , of the HII region be when the internal and the external pressure are matched assuming that the ratio of sound speeds in the ionized gas and the undisturbed gas is 10:1? **Answer:** 201pc

c. Will the star die as a supernova before the HII region reaches its maximum extent? (Assume that the HII region expands rapidly to its initial radius, and then expands at the speed of sound in the ionised gas, $c_{II} = 14$ km s^{-1} to its final radius, and that the age of O stars when they explode is given $\tau = 5(M/40M_\odot)^{-0.4}$ Myr.) **Answer:** yes.

d. As the ionization front advance velocity slows during the evolution of the HII region, a radiative isothermal shock is detatched from the ionisation

front into the undisturbed neutral gas ahead, with sound speed c_I. This shock compresses the gas so that, across the ionization front, the pressure in the neutral gas approximately matches the pressure in the ionized gas. The structure of the HII region in its late evolution therefore consists of a dense thin shell of neutral gas, driven outward by the pressure in the ionized gas, and the sphere of ionized gas in which the ionizations produced by the central star match the recombinations in the ionized gas. All the gas swept up in the expansion of the HII region is contained in one or other of these components. If the density profile of the undisturbed gas can be written as $n(r) = n_0(r/R_0)^{-1}$, and assuming that the HII region reaches its Strömgren radius in a time t_0, what is the equation of motion describing the expansion of the HII region, given in terms of whatever of the following parameters are needed: $S_*, \alpha, n_0, R_0, t_0, c_{II}$ and c_I?

Exercise 9.4.4. An active galaxy is producing an isotropic flux of 10^{53} ionising photons per second, and has aloso produced a radio-emitting jet which has propagated to a distance of 1 kpc from the central source. Assuming that the jet is surrounded by a photoionised medium with a density of 100 cm^{-3}, and that this layer absorbs all of the ionising photons from the central source,

a. At what frequency will the photoionised gas present a free-free optical depth of unity to the radio emission from the jet?

b. Assuming that the jet radio emission has an intrinsic spectrum $F(\nu) = F_0(\nu/1.0\text{GHz})^{-0.5}$ what is the shape of the spectrum we observe at earth?

10. Parameters of Photoionised Regions

"I sit here on the perfect end of a star, watched light pour itself towards me.
The light pours itself through a small hole in the sky.
I'm not very happy, but I can see how things are faraway"
— Richard Brautigan (The Pill versus the Springfield Mine Disaster)

In this chapter we examine how photoionisation theory can be, and has been applied to the derivation of physical parameters and chemical abundances in a wide variety of photoionised plasmas ranging from normal H II regions, plantary nebulae and nova shells to much more exotic objects such as photoionising shocks in metal-rich supernova remnants or active galaxies.

10.1 Nebular Parameters

10.1.1 Nebular Temperature

A fundamental parameter for the derivation of all other physical and chemical quantities in photoionised diffuse plasmas is the electron temperature, since virtually all the observables are strong functions of the temperature.

We have already touched upon the most common means of estimating the electron temperature in Section (3.2); using two emission lines of the same element which have different thresholds for collisional excitation. Generally speaking, in a temperature-sensitive ion with a well-separated triplet of fine-structure terms, we would avoid using the ratio of the transitions from the upper to lower level $(3-1)$ and the middle to the lower level $(2-1)$, since these are usually quite widely separated in wavelength, making the flux calibration and correction for dust absorption rather difficult. Instead, we usually take the $(3-2)/(2-1)$ ratio for which the two emission line tend to occur in roughly the same region of the spectrum. For example, we would use the [O III]$\lambda\lambda$4363/5007Å ratio in preference to the [O III]$\lambda\lambda$2315,21/5007Å ratio. If the two lines are strong enough, then the [O III]$\lambda\lambda$2315,21/4363Å might be used as a secondary reddening estimator, since the two transitions here arise from a common upper level. Examples of some forbidden lines which are commonly used for temperature determinations at optical wavelengths are given in Table (3.1).

The obvious observational problem involved with the use of such ratios is that, when the atomic parameters have been factored in, the ratio of the two transitions depends on the temperature. The $\exp[-E_{23}/kT_e]$ factor for the [O III]$\lambda\lambda$4363/5007Å ratio is $\exp[-32980/T_e]$; *see* Figure (**??**). Thus at normal nebular temperatures ($T_e \sim 10^4$ K), the [O III] λ4363Å line is only a few percent as strong as the λ5007Å line, so that high quality spectra are required to give an adequate signal to noise to measure the fainter line.

A more insidious problem was pointed out by Peimbert (1969) and still has not been resolved to everyone's satisfaction. Normal H II regions are not homogeneous, but contain temperature gradients, dense inclusions in which collisional de-excitation of cooling lines leads to higher temperatures, as well as colliding supersonic flows in which both density and electron temperatures may be raised as a result of shocks. In all such regions of enhanced temperature the emission line flux ratio is raised by the factor given above. This would not be a problem except for the fact that the line emissivity is also raised in such regions by a factor n_e^2. Thus, the temperature estimate provided by the forbidden line ratio is dominated by the line ratio characterising these overdense inclusions, rather than providing a measure of the electron temperature of the nebula as a whole. The electron temperature is therefore systematically overestimated. Photoionsation models by Gruenwald & Viegas (1995) show that this effect can be particularly serious in planetary nebula excited by hot central stars with temperatures in excess of 10^5 K. This can have a serious effect on attempts to determine chemical abundances since the ratio of the strengths of the emission lines of heavy elements to the hydrogen recombination lines is a strong function of temperature. Overestimates of the temperature lead to systematic underestimates of the abundances; see for example Mathis, Torres-Peimbert & Peimbert (1998, and references therein).

A second technique is to use the ratio of a recombination line such as Hβ to the nebular continuum. Since we cannot easily measure the absolute value of the nebular continuum in the presence of contamination from scattered stellar continuum, we measure instead the change in the contiuum across the Balmer series limit (the *Balmer jump*). Since in recombination He^{++} also produces a bound-free continuum with a jump at the Balmer series limit, this must be taken into account. For a given source, if ΔF_{BL} is the measured size of the change in the continuum flux across the Balmer jump (erg cm^{-2} s^{-1} Hz^{-1}), and $F_{H\beta}$ is the Hβ flux from the nebula (erg cm^{-2} s^{-1}), then, to a good approximation:

$$\frac{F_{H\beta}}{\Delta F_{BL}} = 4.498 \times 10^{13} \left(1 + \frac{N(He^{++})}{N(H^+)}\right) T_4^{0.645} \text{Hz} \qquad (10.1)$$

with T_4 being the electron temperature in units of 10^4 K. Thus, at 10^4 K, the difference in continuum flux measured on either side of the Balmer jump extended over 199Å of spectrum would provide a total flux which is just equal the Hβ flux.

This technique has not been applied in a great number of nebulae, but see Hua (1974), since it requires both good spectral resolution and high signal to noise. Futhermore, the temperature sensitivity of the ratio is not very great. Nonetheless, in principal, it provides one of the best and most direct means of measuring the hydrogen recombination temperature that we have.

In principle, very direct means of estimating temperature is simply to measure the brightness temperature of the H II region in the radio continuum at low frequencies where it is optically thick to free-free absorption. In these circumstances (provided that the region is properly resolved by the radio telescope being used), the brightness temperature is equal to the electron temperature. However, the temperature that is being measured here is the temperature of the "photosphere", a thin layer at the nearside of the region, which may not be representative of the volume of the H II region as a whole.

Instead, at radio frequencies, the temperature is usually inferred from the ratio of a radio recombination line to the flux in the adjacent radio continuum. This requires the use of high-frequency data to ensure that the nebula is optically thin and that impact broadening of the line is negligible (*i.e.* it has a Gaussian line profile). With the further approximation that both the line and the continuum are transferred under condition of LTE, the *LTE electron temperature*, T_e^*, can be obtained (Brown, 1988):

$$(T_e^*)^{1.15} = 6350\nu^{1.1}\frac{T_C}{\Delta\nu T_L} \qquad (10.2)$$

where ν is the observing frequency in GHz, and $\Delta\nu$ is the line width in units of km s^{-1}. Because of non-LTE effects, $T_e^* < T_e$ in general, and in practice a full non-LTE analysis is required. The classical study of Galactic H II regions by Shaver *et al.* (1983) gives the appropriate correction:

$$\left(\frac{T_e}{T_e^*}\right)^{1.15} \sim b_n\left\{1 + \frac{1}{T_e}\left(\frac{kT_e}{h\nu}\frac{\Delta b}{b_n} - 1\right)(T_C/2 + T_{BB} + T_{NT})\right\} \qquad (10.3)$$

where T_C is the brighness temperature of the radio continuum from the H II region, T_{BB} is the brightness temperature of the cosmic microwave background, and T_{NT} is the brightness temperature of any non-thermal background source which might be present.

Since the central stars of normal H II regions have similar effective temperatures, then following the discussion of section (9.2.3) it is evident that their electron temperature is then largely determined by the cooling rate, *i.e.* the abundance of the heavy elements. The study of Shaver *et al.* (1983) provided very clear and compelling evidence for the large-scale Galactic abundance gradient simply by measuring the electron temperatures of the H II regions at radio frequencies as a function of Galactocentric distance; (*see* Figure (10.1)).

The final technique for measuring temperature does not yield the electron temperature but rather, the ionic temperature. The quantity that is observed
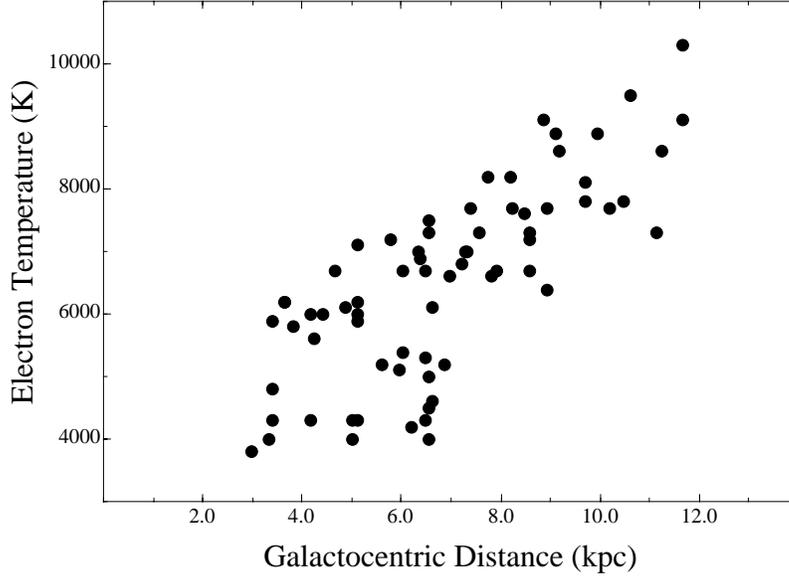
Fig. 10.1. Recombination line temperatures for Galactic H II regions adapted from Shaver *et al.* (1983), adjusted for a solar galactocentric distance of 8.5 kpc. The strong gradient in electron temperature indicates the presence of a galactic logarithmic abundance gradient in oxygen $\sim$0.07 dex kpc^{-1}.

is the thermal Doppler motions along the line of sight. Consider that we measure the full width at half maximum of the line profile, $\Delta\nu_{obs}$, in a line at frequency ν_0 belonging to a particular ion with mass m_i. By themselves, the line-of-sight thermal motions will generate a Gaussian line profile with a shape function $\Phi(v)$ (such that $\int \Phi(v)\,d\nu = 1$) given by (*c.f.* equation (4.2)):

$$\Phi(v) = \left(\frac{m_i c^2}{2\pi k T_i}\right)^{1/2} \exp\left[-\frac{m_i c^2 (\nu - \nu_0)^2}{2 k T_i \nu_0^2}\right] \tag{10.4}$$

and the Doppler line width (full width half maximum) is;

$$\Delta\nu_{therm}^2 = 8\ln 2 \frac{k T_e}{m_i c^2}\nu_0^2 \tag{10.5}$$

Clearly, we require to observe a light element such as hydrogen or helium in order to provide sufficient thermal width. The observed line width is the resultant of this thermal broadening, the Doppler line width associated with turbulent motions, $\Delta\nu_{turb}$, and any line width resulting from unresolved fine-structure or hyperfine line splitting, $\Delta\nu_{FS}$. Assuming that these add in quadrature:

$$\Delta\nu_{therm}^2 = \Delta\nu_{obs}^2 - \Delta\nu_{turb}^2 - \Delta\nu_{FS}^2 \qquad (10.6)$$

In order to derive the temperature therefore, we require an independent determination of $\Delta\nu_{turb}$. This is provided by means of bright forbidden lines (Dopita 1972, 1973), provided that they sample a sufficient fraction of the ionised volume to provide a proper representation of the turbulence in this volume. This technique has been little used at optical frequencies, since it requires both high spectral and spatial resolution, and low turbulence in the H II region to provide sufficiently accurate results.

At radio frequencies, the He and H recombination line profiles may in principal be used to derive the electron temperature, but non-LTE effects must be taken into account. Although the C recombination lines are visible, these cannot be used, since they arise largely in the photodissociation regions beyond the ionisation fronts, and have completely different turbulent, thermal, and collisional widths.

10.1.2 Ionised Masses & Densities

To determine the ionised mass requires that we know the distance to the ionised volume. For extragalactic sources, we rely on the standard methods of distance determination based on an assumed value of the Hubble's constant, with appropriate corrections for the local streaming motions. For Galactic sources, the uncertainties are higher, and depend upon whether the distance is determined from purely dynamical techniques, whether a foreground dust extinction model has been used, or whether the distances of the exciting stars themselves have been independently determined. However, it is not within the scope of this book to review distance determination techniques, so we will assume that the distance is known. In this case, the absolute flux in any process which is driven by recombinations may be used to determine the ionised mass and root mean square density. Take the Hβ line as an example. The observed Hβ flux, corrected for extinction along the line of sight, $F_{H\beta}$, (erg cm^{-2} s^{-1}) is simply the flux produced by all of the recombinations occurring in the nebula diluted by the area over which they are spread at the observed distance of the nebula D:

$$F_{H\beta} = \frac{V}{4\pi D^2} h\nu\alpha_{H\beta}^{eff}(T_e)\, n_{H^+} n_e \qquad (10.7)$$

where V is the ionised volume. The emissivity of Hβ $\varepsilon_{H\beta} = h\nu\alpha_{H\beta}^{eff}(T_e)$ can be approximated in case B and in the temperature regime $T_e \equiv 10^4 T_4$ K by $\varepsilon_{H\beta} \sim 1.235\times10^{-25} T_4^{-0.86}$. Let us take the nebula to be a spherical shell with inner angular radius θ_{in} and outer angular radius θ_{out}, then, taking $n_{H^+} \sim n_H$ and allowing for the electrons produced by ionised helium, equation (10.7) gives the mean surface brightness over the whole nebula (with solid angle θ_{out}^2 sr):

$$F_{H\beta} = 1.27 \times 10^{-4} T_4^{-0.86} \left\langle n_H^2 \right\rangle \left(\theta_{out}^3 - \theta_{in}^3 \right) Y D_{kpc}$$
$$\text{erg cm}^{-2} \text{ s}^{-1} \text{ st}^{-1} \qquad (10.8)$$
$$\text{with: } Y = 1 + \frac{N(He^+)}{N(H)} + 2\frac{N(He^{++})}{N(H)}$$

where D_{kpc} is the nebular distance (kpc). From this equation the root mean square density, $\left\langle n_H^2 \right\rangle^{1/2}$ and the mass of ionised hydrogen:

$$M_H = \frac{4\pi m_H}{3} \left(\theta_{out}^3 - \theta_{in}^3 \right) D^3 \left\langle n_H^2 \right\rangle^{1/2} \qquad (10.9)$$

may be readily obtained.

In many cases, it may be more convenient to measure the free-free flux at radio wavelengths where the nebula is optically thin, rather than the use optical recombination line flux. The radio continuum flux has the advantage that it is unaffected by dust obscuration along the line of sight, or within the ionised volume itself. Provided that the radio telescope has sufficient spatial resolution to measure the nebular angular size accurately, then the radio continuum flux will usually provide a more accurate measurement. The radio continuum flux measures the product $Z^2 n_i n_e$, where Z is the mean charge per ion, rather than the hydrogen density, but since hydrogen produces most of the electrons anyway, these recombination and the free-free continuum fluxes are very closely related. From equation (9.60), we have, including the ionised helium:

$$S(\nu_9) = 4.203 \times 10^5 \nu_9^{-0.1} T_4^{-0.35} \left\langle n_H^2 \right\rangle \left(\theta_{out}^3 - \theta_{in}^3 \right) \left\langle Z^2 \right\rangle Y^2 D_{kpc} \text{ Jy} \quad (10.10)$$

where the flux density is expressed in Janskys (10^{-26} W m^{-2} Hz^{-1}), and the frequency ν_9 is in GHz, and D_{kpc} is the nebular distance (kpc). Note the weaker dependence on the electron temperature here, which means that, in most cases simply taking $T_4 = 1$ would provide an adequate approximation for the purposes of mass estimation. Note that the ratio of equations (10.8) and (10.10) is particularly simple, and is useful for converting between optical and radio flux units:

$$\frac{F_{H\beta}}{S(\nu_9)} = 3.02 \times 10^{-10} \left\langle Z^2 \right\rangle^{-1} Y^{-1} \nu_9^{0.1} T_4^{-0.51} \qquad (10.11)$$

The nebular mass could also (in principle) be estimated from the size, using the density derived from a density-sensitive line ratio such as [O II] $\lambda\lambda 3726/3729$Å or [S II]$\lambda\lambda 6731/6717$Å, as described in Section (3.2.2). The difficulty here is that the local density measured by such ratios is always higher than the root mean square density measured by the nebular flux. This applies to H II regions, planetary nebulae, nova shells and to the narrow line regions (NLR) of active galactic nuclei (AGN). The reason for this is that there exist inclusions of high density gas with high emission measure in

the ionised region. This dense material is either streaming off dense neutral clouds or else is shock-compressed gas produced where supersonic gas streams meet. The ratio of these two densities $n_e / \langle n_e^2 \rangle^{1/2}$ provides an estimate of the *volume filling factor* of the ionised gas, which is an important parameter in photoionisation modelling, since it directly affects the local ionisation parameter and hence the degree of ionisation of the plasma.

10.2 Ionising Source Parameters

10.2.1 UV Flux Distribution

The most obvious way to estimate the source temperature is simply to observe the UV spectral energy distribution (SED) of the exciting source. However, as in life, the most obvious path is not always the best, and there are a number of problems with this approach. Suppose that the central source is a star with a flux distribution approximated by the Black-Body Planck function as given in equation (9.51). The apparent spectral energy distribution is therefore $\nu B(\nu, T)\,\Omega$ (erg cm^{-2} s^{-1}) where Ω is the solid angle subtended by the source at the observer. Most exciting stars of H II regions have spectral energy distributions which peak well into the FUV, in order that they are able to produce an appreciable number of ionising photons. For these, the Rayleigh-Jeans approximation holds in the UV and visible portions of the spectrum; $B(\nu, T) \sim 2\nu^2 kT/c^2$. This is unfortunate, because for hot central sources, the UV flux distribution ($\nu F(\nu) \propto \nu^3$) tells us nothing about the temperature of the central star. The absolute flux might be expected to tell us something about the temperature, but this information cannot be usually be used since we do not, *a priori*, know Ω. If we consider a set of sources having the same luminosity and distance but different temperatures, then from the Stefan-Boltzmann law, $\Omega \propto T^{-4}$. For these, the absolute flux at any particular frequency varies as T^{-3}, making the hotter sources difficult to observe.

In summary, then, the UV flux distribution by itself can only be used to establish the temperature of the central stars when broad spectral coverage and high quality data are available, and where the central star temperature does not exceed about 40000 K. For hotter stars, the UV flux distribution in frequency becomes independent of temperature, and becomes much more difficult to detect against the nebular continuum as the temperature is raised, even if the luminosity remains constant. Figure (10.2) shows one of the rare sucesses of this method, using high quality HST spectra and ground-based data together.

10.2.2 Stellar Atmospheric Modelling

Although stellar atmospheric modelling is not the subject of this book, nontheless we should deal with it briefly, since it provides by far the most accurate and direct means of deriving the stellar parameters. The techniques
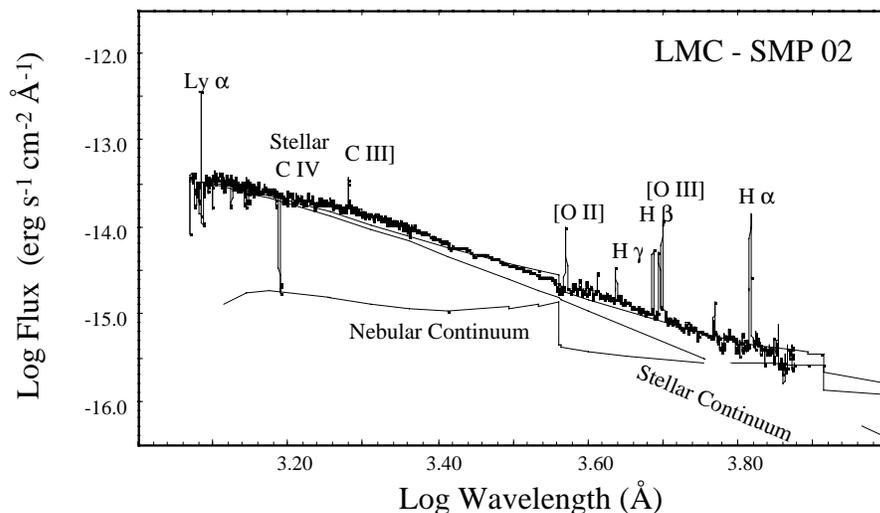
Fig. 10.2. The spectral energy distribution of a low-excitation planetary nebula in the LMC fitted by the theoretical model stellar + nebular continuum (after Dopita *et al.* 1997). Such a fitting only provides useful data about the central star when the effective temperature is less than about 40,000 K.

of detailed stellar atmospheric modelling, including the effects of winds and the spherical extension of the atmospheres, solving the full non-LTE radiative transfer was developed by Gabler *et al.* (1989). In a more restrictive form (plane parallel, hydrostatic) non-LTE models were applied to planetary nebular central stars by Méndez and his co-workers (Méndez *et al.* 1988, and references therein; see also Herrero, Machado & Méndez, 1990). This involves fitting many observed high-resolution line profiles of hydrogen and helium with the theoretical profiles (including both the atmospheric emission and absorption components) to provide a measurement of the effective temperature, T_{eff}, the surface gravity, $\log g$, the helium abundance Y, and the equatorial rotation velocity of the star.

The advantage of locating the central stars on the $\log T_{eff} : \log g$ plane is that this plane can be considered as a distance-independent Hertzsprung-Russell diagram. In addition, the central stars of planetary nebulae (PNe) which Méndez *et al.* (1988) were fitting have unique tracks on this diagram so that, in addition the core mass of the central star can also be derived, and the theoretical flux at some frequency can also be compared with the observed flux to derive the distance of the star to a high degree of precision.

The disadvantage of the stellar atmospheric modelling technique is that it may only be used for nearby and relatively bright objects for which high signal to noise spectrophotometry is available. However, the new generation of 8-10 metre class telescopes at sites of good seeing now enables such analysis

to be applied to the very large population of PNe in the Galactic bulge and to the hot OB stars in the LMC.

10.2.3 Zanstra Temperatures

In Section (10.2.1) we showed how, by itself, the UV spectral energy distribution is not a very useful technique in estimating the stellar temperature. The core of the problem is that the observations do not extend to short enough wavelength. What we need is a means of probing into the EUV. Fortunately, we have a means to do this. The nebula itself absorbs the A-band to C-band ionising photons, producing H^+, He^+ and He^{++} zones in which H I, He I and He II recombination lines are produced. Thus, provided that the nebula can absorb these photons (that is to say, that the nebula is optically thick to the escape of the ionising radiation field), then the flux in the H I, He I and He II recombination lines provides a direct estimate of the fluxes in the band-A to band-C ionising radiation field. If furthermore we measure the flux produced in the stellar continuum at the same wavelength as the recombination lines, and form the line to continuum flux ratio, we have a reddening and distance-independent estimate of the shape of the ionising spectrum of the central star.

This then is the essence of the Zanstra (1931) technique of measuring effective temperatures, which its inventor was fond of describing as "a cheap way to do space science", since through observations in the optical we can essentially accomplish EUV astronomy from the ground.

Now to the details. The $H\beta$ flux has already been given in equation (10.7). Using this to work out the total number of recombinations, and matching these to the number of ionisations produced by the central star, S_* :

$$F_{H\beta} = \frac{V}{4\pi D^2} h\nu_{H\beta} \alpha_{H\beta}^{eff}(T_e) n_{H^+} n_e = \frac{S_*}{4\pi D^2} h\nu_{H\beta} \left(\frac{\alpha_{H\beta}^{eff}(T_e)}{\alpha_H^B(T_e)} \right) \quad (10.12)$$

where $\alpha_H^B(T_e)$ is the Case B recombination coefficient for hydrogen. However, for a Black-Body distribution:

$$S_* = 4\pi R_*^2 \int_{\nu_0}^{\infty} \frac{B(\nu, T)}{h\nu} d\nu \sim 4\pi R_*^2 \frac{\sigma T_*^3}{3k} \quad (10.13)$$

where for the approximation we have used the Stefan-Boltzmann law (with constant σ), and have approximated the mean frequency to the peak of the Black-Body flux distribution given by Wein's displacement law. However, the observed stellar flux at $H\beta$ is:

$$F_* = \frac{R_*^2}{D^2} B(\nu_{H\beta}, T) \sim \frac{R_*^2}{D^2} \frac{2\nu_{H\beta}^2 k T_*}{c^2} \quad (10.14)$$

where the approximation uses the Rayleigh-Jeans law.

Combining equations (10.12) to (10.14) we have:

$$T_*^2 = \frac{6k^2\nu_{H\beta}}{hc^2\sigma}\left(\frac{F_{H\beta}}{F_*}\right)\left(\frac{\alpha_H^B}{\alpha_{H\beta}^{eff}}\right) \tag{10.15}$$

From this it follows, to a first approximation, in an optically-thick nebula the equivalent width of $H\beta$ measured with respect to the stellar continuum increases as the square of the stellar temperature. Similar expressions can be derived for the He I and He II equivalent width.

In applying the Zanstra method, we do not use approximations such as (10.15), which are useful only insofar as they illustrate the basic physics of the technique. Nor do we usually measure F_* at the wavelength of the recombination line, but instead we measure the V magnitude. Gabler *et al.* (1992) give the following useful relationships for the hydrogen and helium *Zanstra Ratios*, Z_H and $Z_{\text{He II}}$, defined as the logarithm of the ratio of the number of ionising photons $(\text{cm}^{-2}\text{s}^{-1})$ to the stellar continuum at V (erg cm^{-2} s^{-1} Hz^{-1}). For hydrogen the number of ionising photons is the sum for $v > v_0$, while for He II the is the sum over $v > 4v_0$:

$$Z_H \quad = 30.825 + \log F_{H\beta} + 0.4V + 0.13c + \log\left(\alpha_H^B/\alpha_{H\beta}^{eff}\right) \tag{10.16}$$

$$Z_{\text{He II}} = 30.809 + \log F_{4686} + 0.4V + 0.17c + \log\left(\alpha_{He}^B/\alpha_{4686}^{eff}\right) \tag{10.17}$$

where c is the logarithmic dust extinction at $H\beta$ (see chapter on dust, below). The value of the Zanstra ratio is a strong function of stellar temperature, and is best determined from the stellar atmospheric theory.

If we approximate the stellar energy distribution by a Black-Body function, as was common a few years ago, then for planetary nebulae the Zanstra temperature determined from the He II lines, T_Z (He II) frequently turned out to be much higher than the hydrogen Zanstra tempertature, T_Z (H). This is the famous *Zanstra discrepancy*; see, for example, Kaler (1989). There are two possible explanations for this, either the nebula is failing to absorb all the H-ionising photons because it is optically-thin, or else the assumption of a Black-Body distribution is wrong. While the first explanation is undoubtedly correct in some cases, the second one must be invoked in others. Initially, the continuum blanketing of LTE model atmospheres made the Zanstra discrepancy even worse, and this was not improved by non-LTE atmospheres based on the plane-parallel and hydrostatic assumptions. Because the He II ionising photons are produced high in the atmosphere, the "unified" non-LTE models including spherical extension and stellar winds improved, but did not totally eliminate the problem (Gabler, Kudristzki & Méndez, 1991). The likely explanation is that the shocks in the stellar wind are producing the additional hard UV photons which are required. As a consequence, the best compromise

at present appears to be to use the "unified" non-LTE models to describe the flux distribution down to the He II limit $\nu < 4\nu_0$, and then simply to assume a Black-Body distribution at shorter wavelengths, with an effective temperture equal to the effective temperature given by the non-LTE model.

The major observational difficulty in the application of the Zanstra technique is in measuring the optical flux from the central star. As equation (10.15) shows, the ratio of the Hβ and the continuum flux varies as T_*^2, approximately. However, the nebular flux near Hβ is mostly hydrogenic free-bound continuum with some two-photon contribution, and therefore scales with the Hβ, approximately. Inevitably, therefore, as the temperature of the central star increases, its optical flux becomes more and more difficult to detect against the nebular background. This problem increases with increasing distance, since the nebular surface brightness remains constant for as long as it is resolved, while the flux from the central star decreases as the square of the distance. In addition, we know that internal obscuration due to dust can be important in both H II regions around OB stars, or within planetary nebulae shells. If we were unlucky, the central star might be obscured, or partially obscured, by such a dust cloud, which would completely invalidate attempts to derive Zanstra temperatures in these cases.

A different problem is that the nebula has to absorb all of the photons emitted by the central star in order for the count of recombinations to provide a proper measure of the number of ionising photons. That is to say, the nebula must be *optically thick* to the ionising radiation in all directions. This requirement may be relaxed somewhat if we are able to construct a self-consistent photoionisation model for the nebular, which would provide a theoretical estimate for the photon leakage in the various ionising wavebands.

10.2.4 Energy Balance (Stoy) Temperatures

The thermal balance condition for the ionised nebula was given in equation (9.2);

$$\dot{Q} = \int_{\nu_0}^{\nu_{\max}} \frac{(\nu - \nu_0)}{\nu} I(\nu)\,\kappa_\nu d\nu = \Lambda(T_e)\,n n_e \qquad (10.18)$$

It is obvious, on inspection of this equation, that the "harder" the radiation field, then (for a given opacity) the more energy is delivered per photoionisation, and this has to be matched by an increase in the cooling rate $\Lambda(T_e)$, produced by the sum of all the collisionally excited lines, continuum processes *etc.* Note that this equation implies that, for a given set of chemical abundances, nebulae excited by hotter stars have to have higher electron temperatures in order to provide the greater cooling rate. This is consistent with what we learnt in section (9.2.3), and conforms to what we observe in practice.

If the ionised volume is V and the distance of the nebula D then the observed flux (erg cm^{-3} s^{-1}) in all the cooling processes is simply $\dot{q} = V\,\dot{Q}$

$/4\pi D^2$. However, in equilibrium, the number of ionising photons is matched by the number of recombinations, and this quantity can be measured by the flux in a recombination line such as Hβ. Thus, if we form the ratio $\dot{q}$ $/F_{H\beta}$, this in essence measures the mean energy delivered by the photon field per photoionisation, which is directly related to the effective source temperature. Taking this idea a little futher, the observed Hβ flux is given by equation (10.12), while $\dot{q}$ is:

$$\dot{q} = \frac{S_* h \left(\langle \nu \rangle - \nu_I \right)}{4\pi D^2} \tag{10.19}$$

where $\langle \nu \rangle$ is the energy weighted mean frequency of the ionising field ($\sim 3kT_*/h$ according to the Wien law for a Black Body), and ν_I is the effective mean ionisation potential of the ions which are absorbing this field (mostly H and He in normal H II regions), The effective mean ionisation potential here is slightly larger than the mean ionisation potential appropriate to the ionisation state of the plasma, since the emitted photoelectrons must have an energy greater than the mean thermal energy of the particles in the plasma to provide a net heating effect. It follows therefore that:

$$\frac{\dot{q}}{F_{H\beta}} = \left(\frac{\alpha_H^B}{\alpha_{H\beta}^{eff}} \right) \frac{\left(\langle \nu \rangle - \nu_I \right)}{\nu_{H\beta}}$$

This is the physical essence of the Stoy (1933) energy balance technique, generalised by Kaler (1976) and Preite-Martinez & Pottasch (1983). The reader is referred to the Preite-Martinez & Pottasch paper where convenient expressions for the application of the technique are given, for stars with both Black-Body distributions, and for various model atmospheres applicable to the central stars of planetary nebulae.

The Stoy energy balance technique provides a measurement of the colour temperature of the exciting star or stars, and can be applied, with the appropriate expressions developed by Preite-Martinez & Pottasch, even to nebulae which are optically-thin. It has the further advantage that only nebular data are required, which means that, unlike the Zanstra method, it can be applied to distant objects with hot central stars.

The major difficulty with the method is that we need to know the sum of *all* of the collisionally-excited lines in the nebular spectrum, but only rarely do we have the luxury of the multi-waveband data that would make this measurement possible. Instead we have to develop techniques of correcting for unobserved cooling lines using data obtained in a much more restricted waveband; usually only the visible spectrum is used for this purpose.

10.3 Photoionisation Modelling

The fundamental purpose in constructing a photoionisation model for a particular object is not only to gain insight into the nature of the exciting source

and the physical conditions in the surrounding nebula, but also to be able to quantitatively determine the chemical abundances which characterise that nebula. In "normal" H II regions, these tell us about the chemical composition of the interstellar medium as a function of position in a galaxy, and of type of galaxy, and so lead us to an insight as to the way galaxies form and evolve. In planetary nebulae, mass-loss bubbles, nova shells and photoionised portions of supernova remnants, we learn about the specific processes of nucleosynthesis which have occurred within individual stars prior to this nuclear-burnt material being ejected back into the interstellar medium.

However, the number of photoionisation models which it is possible to construct is infinite. This is true if we restrict ourselves to only considering spherically-symmetric models. Even in choosing the spectral energy distribution of the exciting star, we have an embarrasment of choice. Do we use a simple Black-Body distribution or a model atmosphere. If the latter, then which one? In addition, we must also choose the luminosity of the source. For the nebula we have the density, the density distribution, the nature and degree of clumping of matter and the geometry of the ionised gas with respect to the ionising source. Finally, we have to choose the chemical abundance set for as many as fifteen elements which are either significant or else trace coolants within the gas. This leads to as many as twenty five parameters which have to be fitted to the data.

In the light of this, it might be considered remarkable that any progress has been made. However, the theory of photoionised regions tells us how to reduce the free parameters to a more manageable number. First, we should note that the *spectral energy distribution (SED)* of the central source is vital. It largely determines the gross ionisation structure in hydrogen and helium, to which the ionisation structure of the other elements is coupled. However, the spectral energy distribution is not sufficient to determine the ionisation structure. For example, a powerful source with a low temperature will produce a low-excitation nebula. However, a weak source with a high temperature will also produce a low-excitation nebula, albeit of a subtly different kind. In order to determine the excitation, we need to know, in addition to the SED of the central source, the *mean ionisation parameter* $(\mathcal{U})$ in the ionised gas.

Provided that the nebula is optically-thick to the ionising radiation, it turns out that, for a given abundance set, the ionisation parameter $\mathcal{U}$ and the SED are sufficient to define an almost unique emission spectrum, regardless of the detailed geometry of the gas with respect to the central source. For example the gas could be in a uniform shell of large radius, or else clumped in dense blobs within a smaller radius. Provided that the average hydrogen atom sees the same mean ionisation parameter, the degree of ionisation and emitted spectrum is the same. Of course, collisional de-excitation in denser plasmas may change the spectrum somewhat by raising the local temperature, or by collisionally suppressing some lines.

The fact that objects with harder SEDs have higher forbidden line to Balmer line ratios (section (10.2.4), above), and that the excitation of the nebula is controlled by both $\mathcal{U}$ and the SED suggests that by plotting pairs of emission line ratios, in particular, the ratio of a forbidden to a Balmer line against an excitation-dependent ratio such as [O II]/[O III] we will be able to separate different classes of photoionised nebulae. The utility of such *diagnostic diagrams* has been amply demonstrated over the years, and we will have more to say on the subject below. Baldwin, Phillips and Terlevich (1981) were the first to demonstrate how this procedure could be used to separate the various classes of active galactic nuclei. However, such diagrams had been used for many years before this in the separation of H II regions from shock-excited plasmas, or for abundance diagnostics of emission nebulae (*i.e.* Sabbadin, Minello and Bianchini, 1977, D'Odorico 1978, Dopita 1977, 1978). Veilleux and Osterbrock (1987) and Osterbrock, Tran & Veilleux (1992) emphasised the importance of choosing line ratios which are as close together in wavelength as possible to avoid errors due to reddening corrections or errors in instrumental sensitivity calibrations.

10.3.1 Ionisation Correction Factors

In sections (9.2.1) and (9.2.2) we showed how the ionisation balance in the nebula is controlled by the hydrogen and helium, while the ionisation state of the heavier elements is closely locked to that of hydrogen and helium. This division of the nebula into a set of "onion-skins", each with a different (and well-defined) ionisation state is the basis for the *ionisation correction factor (ICF)* method (Peimbert & Costero, 1969). This enables us to determine nebular abundances, while avoiding the need to construct a particular photoionisation model to fit the spectrum of each object and at the same time attempting to take account of the un-observed ionisation stages of particular elements. This procedure becomes necessary when a nebula is observed in only, say, the optical lines it produces. In this case, the representation of the various ionisation stages of the different elements in the spectrum is rather patchy, as is the information that can be gleaned from a high signal to noise spectrum.

For example, oxygen is particularly well catered for. Atomic oxygen can be seen in the [O I]$\lambda\lambda$6300,6363Ålines or the [O I]λ5577Ålines, and the ratio of these is an excellent temperature diagnostic in most cases. Singly ionised oxygen is observed through the [O II]$\lambda\lambda$3727,3729Å(which can be used to determine the electron density, *see* Chapter 3) and the red [O II]$\lambda\lambda$7318,7328Ålines, which when ratioed with the near-UV lines provide a temperature diagnostic. Doubly ionised oxygen is represented by the very intense [O III]$\lambda\lambda$4959,5007Ålines (which are often the principle coolant in photoionised nebulae), and by the [O III]λ4363Åline, which ratioed with either or both of the other lines provides a temperature determination. Thus, for oxy-

gen, both the ionisation structure and the physical parameters of the nebula are rather well determined.

For the other elements, the situation is usually more difficult. For example, atomic nitrogen is seen in the [N I]$\lambda\lambda$5198,5200Ålines (which provide a density diagnostic). The lines that could provide the temperature diagnostic lie too far into the IR to be observed in most cases. Singly ionised nitrogen is also seen in the red [N II]$\lambda\lambda$6548,6584Ålines and the [N II]λ5755Åline, which provides a temperature diagnostic for the singly ionised zone. All the other useful lines of these ionisation stages and of others are in the space IR or UV region. The case of carbon is even worse. Here we only have the [C I] $\lambda\lambda$9823,9850Å, and very weak recombination lines in the visible, all the other strong lines being found at the vacuum UV wavelengths (e.g. C II λ2326Å, C III] $\lambda\lambda$1907,9Å, or the C IV$\lambda\lambda$1548,1550Ålines).

In the ICF method, we would attempt to construct empirical relationships based upon consideration of the ionisation potentials, or semi-empirical relationships based on more detailed photoionisation modelling to account for the missing ionisation stages. For example a reasonable approximation for nitrogen (although not necessarily the best one) in a nebula excited by an O-type star would be to assume it has the same ionisation structure as the oxygen ions. This would then allow us to determine the total nitrogen abundance from only the [N II] lines, with the help of the temperatures determined in the other zones using [O I] and [O III] line ratios:

$$\frac{N(N)}{N(H)} \sim \frac{N(N^+)}{N(H^+)} \left(\frac{N(O^0) + N(O^+) + N(O^{++})}{N(O^+)} \right)$$

More sophisticated ICF schemes have been developed on the basis of detailed photoionisation models (e.g. Peimbert & Torres-Peimbert, 1977; French & Grandi, 1981; Stasinska 1978, 1981). Whilst these are still useful for obtaining approximate abundances and physical conditions, the ready availability of codes such as CLOUDY or MAPPINGS has made such procedures somewhat obsolete today. The 3-D modelling code of Gruenwald *et al.* (1997) has recently been applied to the analysis of the nature and the size of the error that such ICF methods would make when applied to observations which do not integrate the line flux over the whole nebula, such as would be the case for a long-slit spectrum of a resolved nebula. In some cases, these errors can be large (Gruenwald & Viegas, 1998).

10.3.2 Self-Consistent Photoionisation Modelling

The availability of high-quality, broad waveband spectroscopy and imaging is now making it possible to produce self-consistent photoionisation models from which all the physical parameters can be derived. A self-consistent model ideally uses a multi-zone approach to simulate the three-dimensional structure, which is particularly important for modelling bipolar structures. For

example, Dopita *et al.* (1997, and references therein) use two-zone model for planetary nebula modelling. This consists an isobaric, optically-thick model to reproduce the dense equatorial ring where the ionisation fronts are located, and a power-law density distribution with a lower inner density to simulate the gas streaming away from these into the polar lobes. This distribution may be either optically-thin or optically-thick. Recently, continuous 3-D photoionisation models have now been produced and applied to real objects by Gruenwald *et al.* (1997).

The goal of self-consistent photoionisation models is to match the observed linear size and, if possible the 3-D morphology. The model must also provide the same line-of-sight densities as observed from [O II], [S II] or the UV lines, as well as successfully reproducing the reddening corrected Hβ flux or the radio continuum flux from the nebula. The stellar and nebular parameters and chemical abundances of each element are adjusted until the dispersion of the observed *vs.* the theoretical line intensities is minimised for all ionisation stages of each element. In this procedure, the overall degree of excitation is largely determined by the assumed effective temperature of the central star.

The optical depth to the ionising continuum flux is an important parameter, and some lines prove to be quite sensitive to this parameter. In particular, The absence of an outer nebular zone is signalled by particularly weak low excitation lines such as C II], [N I], [N II], [O I] and [O II], unusually high electron temperature, or else particularly strong high excitation lines such as He II, [N V], [Ne V] and [Ne IV]. All of these are symptomatic of an optically-thin nebula. In this case, the effective temperature of the central source together with the optical depth is defined by the excitation of H and He, combined with such lines or line ratios which are sensitive to the optical depth. In the case of an optically-thick nebula, the excitation state alone is sufficient to define the effective temperature.

For cooler objects the central star is directly visible through its UV continuum. A comparison of the predicted nebular+ stellar continuum with that which is observed always provides a "sanity check" of the model, and in some cases may also allow us to obtain an independent estimate of the luminosity of the central star, provided that we can assume that the temperature of the central star has been accurately determined from the nebular model. An attempt to directly determine the temperature from a variant of the classical Zanstra method would often be invalidated in the optical thin cases.

Finally, the abundances can be determined by adjusting them until the RMS scatter of the observed line strengths of each element with respect to, say, Hβ is minimised. In this way, a formal error of about 10% in the chemical abundances of most elements can be achieved. Real errors may be somewhat larger, limited by the quality of the atomic data. If the modelling has been done correctly, then the electron temperature indicated by temperature sensitive line ratios should agree with observation, within differences that can

be accounted for by the (unmodelled) temperature fluctuations in the ionised plasma.

10.4 Abundances

All of the observational techniques described in the previous sections of this chapter can be, and have been, applied to the analysis of observations of the common classes of photoionised nebulae. A knowledge of the physical and structural parameters of such nebulae is a necessary prerequisite to the derivation of the chemical abundances which characterise them. In turn, these chemical abundances can be used to infer details about the thermonuclear reactions which occurred in the stars which gave rise to the nebula, or, in the case of H II regions, to discover details of the chemical evolution of the interstellar medium in the host galaxy. Here we will briefly summarise some of the results of this work for three major classes of photoionised nebulae.

10.4.1 Galactic Abundance Gradients from H II Regions

When a hot, young star photoionises its placental cloud of interstellar gas, it provides us with a means of studying the chemical composition of this material. Because they are bright and their spectra can be readily analysed, H II regions have been extensively used to study abundance variations in the interstellar media of external galaxies. Early photographic spectra showed that, in disk galaxies, the H II regions in the outer spiral arms displayed large [O III]/Hβ ratios, while in the inner regions the [N II]/Hα ratios were relatively large (*i.e.* Searle 1971). It was soon realised (*e.g.* Sanduleak, 1969) that this was due to a global abundance gradient, which results from many generations of stars gradually depleting the gas content of the interstellar medium, and at the same time chemically polluting it by mixing their nucleosynthetic products back at the end of their lives. Because this star-gas cycle works more completely in the inner regions of the galaxy, the abundances of the heavy elements relative to hydrogen are higher there. A number of factors tend to alter the size of these gradients. Mass infall may be an important parameter. Where strong stellar bars exist, radial mixing tends to flatten these gradients, and in low mass systems, global mixing also reduces or eliminates the gradient. In low mass systems, loss of the nucleosynthetic products in galactic winds may also be important. Thus, both the extent and the magnitude of these gradients provides observational constraints on models of disk evolution. For recent discussions, see Matteucci & Francois (1989), Götz & Köppen (1992), Prantzos & Aubert (1995) and Friedli & Benz (1995).

The evidence of a gradient in our own Galaxy derived from radio recombination line temperatures has already been mentioned. In fact, a number of other techniques involving different classes of object have been used to

establish the abundance gradient in oxygen, and these are now in quite good agreement one with another. There are shown in Table (10.4.1). The logarithmic oxygen abundance gradient $d[O/H]/dr$ in the Galaxy is about -0.07 dex kpc^{-1}.

Objects Used	Technique	Gradient (dex kpc^{-1})	Ref
H II regions	Recomb. line temperatures & optical spectra	-0.07±0.02	1
H II regions	Recomb. line temperatures ultracompact H II regions	-0.05±0.01	2
H II regions	far-IR fine structure lines	-0.08±0.01	3
H II regions	far-IR fine structure lines: compact H II regions	-0.064±0.009	4
PNe	Optical spectra of PNe: photoionisation analysis	-0.06±0.01	5
Old SNR	Optical spectra of SNR: radiative shock analysis	-0.07±0.02	6
B-Type stars	Stellar spectra: non-LTE analysis	-0.07±0.01	7

[1]Shaver et al. (1983) [2]Afflerbach et al. (1996) [3]Simpson et al. (1995), [4]Afflerbach, Churchwell & Werner (1997) [5]Maciel & Köppen (1994) [6]Binette et al. (1982) (for S rather than O) [7]Smartt & Rolleston (1997).

In an H II region, the principal coolant is oxygen. Curiously enough, the absolute oxygen abundance derived for H II regions in the solar vicinity is consistently lower than that measured in the sun. This is contrary to what would be expected according to theories of chemical evolution. Since the sun formed some 4 Gyr ago it would be expected to have been formed from a less chemically-evolved interstellar medium. Depending on the importance of temperature fluctuations, the O/H abundance of H II regions in the solar neighbourhood is N(O)/N(H) = 4 - 6×10^{-4} (Peimbert, Storey & Torres-Peimbert, 1993) compared with 8.5×10^{-4} for the Sun (Anders & Grevesse, 1989). This result is not an artifact of the method. Chemical analysis of the B-stars in the solar neighbourhood give similar results; N(O)/N(H) = 4.5×10^{-4} (*references in* Meyer, Jura & Cardelli, 1998). High resolution absorption line studies of the local ISM using the GRHS instrument of the Hubble Space Telescope yield a gas-phase abundance of 3.2×10^{-4}. When corrected for the fraction of oxygen locked up in dust grains, the oxygen to hydrogen ratio cannot much exceed N(O)/N(H) = 5×10^{-4}(Meyer, Jura & Cardelli, 1998).

This 'deficit' problem is not confined to oxygen. Similar deficits are seen for the noble gases such as Ne, Ar and Kr, which cannot be caught up in the grains. We have to conclude therefore, that the sun was either born in a region of enhanced metallicity, or that it was born nearer the centre of the Galaxy, and has migrated to its present position by orbital diffusion (Weilen, Fuchs & Dettbarn, 1996).
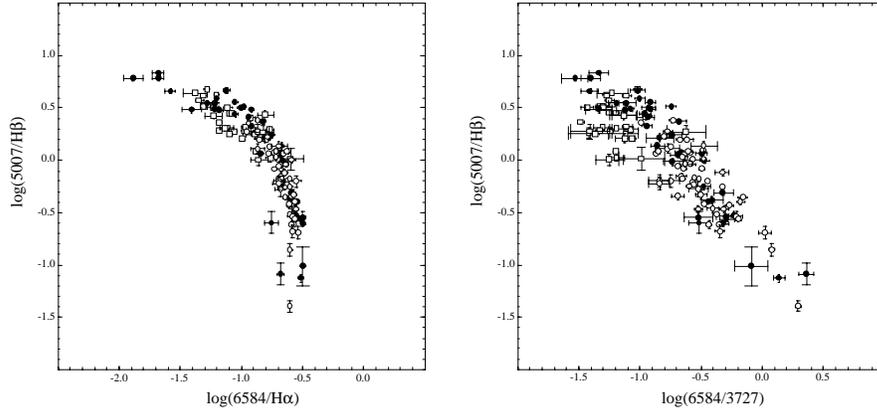
Fig. 10.3. The abundance sequence for extragalactic H II regions defined by observations of M101, NGC1365 and NGC1313. The scatter in the right hand plot better reflects the intrinsic scatter between H II regions, since on the left, variations in stellar temperature do not change the positions of the points, while variations in $\mathcal{U}$ tend to move points along the curve. The horizontal axis in the right-hand plot is sensitive to temperature (data from Kennicutt & Garnett 1996, Roy and Walsh, 1997, and Walsh & Roy 1997).

Turning now to extragalactic H II regions. The primary change in the structure of an H II region as the abundance is increased is that the electron temperature becomes lower. This drives the principal changes in the optical spectra. For example, consider a principal such as O III. As the abundance increases and the temperature decreases, the balance of cooling in the optical forbidden lines [O III] $\lambda\lambda$4959,5007Å is shifted in favour of the far-IR fine structure lines [O III]$\lambda\lambda$52μm, 88μm. Likewise the ratio of temperature sensitive lines such as [O II]$\lambda\lambda$3727,9Å/7318,28Å decreases as the rate of collisional excitation to the more excited state decreases. In fact, the decrease in electron temperature makes many of the traditional temperature-sensitive line ratios very difficult or impossible to measure. However, other pairs of line ratios may be substituted provided that the ratios of the abundances of different elements change in a systematic way with the oxygen abundance. For example, the [N II]/[O II] ratio forms such an abundance-dependent (temperature sensitive) diagnostic (Garnett & Shields, 1987) since the theshold energy required to excite the [O II] lies is so much higher than the threshold for excitation of the [N II] levels.

The result of this is that there is a well-defined extragalactic abundance sequence when H II region data is plotted on various diagnostic diagrams. We show two of these in Figure (10.3).

The calibration of the extragalactic abundance sequence in terms of absolute abundances is the central goal of the analysis. Pagel *et al.* (1979) used an ionisation correction factor argument to suggest that the sum of the bright

forbidden oxygen lines, $R_{23} = ($ [O III] $+$ [O II] $)/$Hβ should represent a good abundance sensitive ratio. Calibrations of this have been proposed by Edmunds & Pagel (1984), Mc Call *et al.* (1985) and Dopita & Evans (1986). However, R_{23} is not the best ratio to use, because it becomes insensitive at the high abundance end as a result of the fading of the [O II] lines at low electron temperatures. Instead, temperature sensitive ratios such as [N II]/[O II] or [S II]/[O II] probably offer better abundance calibrations. For a recent discussion, see Kennicutt & Garnett (1996).

10.4.2 Abundances in Planetary Nebulae (PNe)

Advanced Topic

Planetary nebulae (PNe) acquired their peculiar name because their bright disks of ionised gas, typically a few seconds of arc across, were often mistaken for new planets by early telescope users. They are derived from low-mass stars, below 6 - 8 $M_\odot$, but more typically having masses in the range 1 - 3 $M_\odot$. Such stars evolve to the PNe stage at the end of their nuclear-burning lifetime, when the unburnt envelope mass becomes so low that it reveals the hot core of the star that remains. At this point, the star is either burning hydrogen or helium under electron-degenerate conditions in a thin shell, and the core, supported electron degeneracy pressure, will later evolve to become a white dwarf. Because of their low mass, the lifetime of the precursor star can be a considerable fraction of the age of a galaxy. From the Vassiliadis and Wood (1993) models:

$$\tau \sim 11.0 \, (m/M_\odot)^{-3.1} + 0.46 \, (m/M_\odot)^{-4.6} \, \text{Gyr} \qquad (10.20)$$

where the initial stellar mass is m.

The ionised shell of gas in the PNe phase was thrown off from the central star during a rapid phase of mass-loss which occurs during the asymptotic giant branch (AGB) phase of evolution. This mass-loss is driven by the large radiation pressure in the AGB phase. It is also assisted by the shell flashes which result from burning helium under electron-degenerate conditions. As a result, the mass-loss rate is a very steep function of mass. This serves to channel a wide range of initial stellar masses (1 - $8M_\odot$) into a rather narrow range of final (White Dwarf) stellar mass (0.55 - $1.4M_\odot$). According to the mass-loss formulation adopted by Vassiliadis and Wood (1993) and Marigo, Bressan and Chiosi (1996) the relationship between initial mass, m, and final core mass, m_{core}, can be expressed as:

$$m_{core}/M_\odot = 0.524 + 0.0438 \, (m/M_\odot) + 0.0095 \, (m/M_\odot)^2 \qquad (10.21)$$

this formula agrees with recent observational data derived from observations of white dwarf stars in clusters (Jeffries, 1997).

These basic facts of stellar evolution theory immediately suggest a reason to measure abundances in PNe. Since the central star does not proceed beyond helium burning, and since much of the unburnt envelope is lost, the PNe provide a time-capsule of ionised material which has been chemically unchanged since the star formed, at least as far as the heavy elements are concerned. In principle, PNe may be used to probe the history of chemical enrichment in galaxies, although so far this has done very extensively (but see Dopita *et al.* 1996).

For the lighter elements, and the heaviest elements, the material ejected as a PN shell is not entirely pristine. A variety of convective phases dredge up partially hydrogen-burnt material from the boundary of the core during both the giant and asymptotic giant phases of evolution (Iben and Renzini, 1983; Renzini and Voli 1981). These processes are dependent upon both the mass and initial metallicity of the star and may be summarised as follows:

- The first dredge-up, operating as the star becomes a red giant for the first time, is produced by the penetration of the convective envelope into regions which are partially CNO-burnt. The dredged-up material is mixed throughout the envelope, with an enhancement of the ^{13}C and ^{14}N, and a decrease in ^{12}C abundance.
- The second dredge-up appears in the early AGB evolution of stars more massive than $m > 3 - 5M_\odot$ when the hydrogen-burning shell extinguishes. Once again the base of the convective envelope dips into burnt material. This time envelope enhancements of ^{4}He, ^{14}N and ^{13}C are produced.
- The third dredge-up occurs in the thermally-pulsing AGB phase where, after each He-burning shell flash, the convective envelope dips down, dredging up nuclear processed material rich in ^{4}He, ^{12}C and the $s-$process elements.
- Finally, the so-called *hot-bottom burning* occurs in the more massive AGB stars $(m > 3M_\odot)$ when convection in the stellar envelope cycles matter through the hydrogen-burning shell during the inter-pulse phase, with resultant partial CNO-cycling of the whole envelope. Significant ^{14}N, and possibly ^{4}He production, may occur as a result.

Through these processes PNe affect the chemical evolution of the galaxy, since they are significant sources of He, C, N and the heavy $s-$process elements. At the same time they return an appreciable mass fraction of almost pristine low-metallicity gas back into the interstellar medium, ready to be re-cycled to new stars.

To some extent, all the effects predicted by theory are observed in PNe. However, the quantitative agreement is often poor, and it is clear that this is an area where close contact between theoreticians and observationalists could prove mutually profitable.

Let us first consider what has been learnt in a global sense by comparing abundances of PNe in different galactic environments. At present, we can compare several groups of PNe which are derived from stellar populations with radically different star formation histories.

In the Galactic halo are found PNe derived from stars formed during the initial collapse of the galaxy. At that time the chemical enrichment of the infalling gas was dominated by the massive young stars, which means that the relative abundance of oxygen and the $\alpha-$ process elements compared to iron was higher than it is today in the galactic disk. Since the majority of the infall occurred 10 - 12 Gyr ago, the stars evolving today to the PNe phase had initial masses of order $1M_\odot$.

PNe formed in the Dwarf Elliptical galaxies of the Local Group (and which have been found in the Sagittarius dE and the Fornax dE) have a somewhat different history. Here there are both old and intermediate age stellar populations. As a result, iron, which is formed later in Type I supernova explosions, has a higher abundance than in the Galactic halo. However, at some point, probably as a result of the energy input of these explosions, all of the gas was stripped from these systems and star formation ceased. The progenitor stars of the PNe seen today had masses in the range $1 - 1.5M_\odot$, roughly.

The solar neighbourhood has had a long and rich star-formation history, with stars being formed throughout at least the last 10 Gyr. As a consequence, the PNe in the galactic neighbourhood are derived from stars of up to $6 - 8M_\odot$, although the majority come from a population having initial masses in the range $1 - 2M_\odot$. The most massive population of PNe shows the filamentary bipolar morphology and nitrogen enhancements which characterise the Peimbert Type I objects.

The Galactic bulge PN observed most extensively by Ratag *et al.* (1997) come from such a wide range of intitial masses and metallicities that they cannot be treated as a single population, nonetheless they provide a record of the chemical evolution of the bulge and as such deserve to be studied in their own right.

Finally, the Magellanic Clouds present us with a population in which most of the star formation has occurred relatively recently. In the LMC, the phase of rapid star fomation started about 5 Gyr ago, although the oldest stellar populations have ages similar to the oldest stars in the Galaxy. Currently, most of the PNe are derived from stars with masses in the range $1.5-2.5M_\odot$, although there is an important population of more massive nitrogen-rich Type I objects. HST images reveal that these have bipolar morphologies, just like their galactic counterparts. Chemical evolution of the LMC has proceeded less far than in the galaxy, so that the current ISM abundances are about 0.3 dex lower than in the solar vicinity.

The SMC is similar to the LMC, but is even more extreme. The phase of rapid star formation set in 2 - 4 Gyr ago, star formation is very active today, and the youngest PN represent an even more important component. The base metallicity remains much lower, about 0.6-0.7 dex lower than the solar vicinity.

The results of all of this are evident in Table (10.4.2) which presents the average abundances derived for each of these populations of PNe along with

averages for the ISM, where it still exists. The abundances are expressed in the usual way, as $12 + \log[N(A)/N(H)]$, where A is the element being considered.

Population:	He	C	N	O	Ne	S	Ar
Halo PN[1]	11.00	7.85	7.51	7.98	6.44	6.06	5.17
dElliptical PN[2]	11.04		7.36	8.33	7.56	6.39	5.85
SMC PN[3,4]	11.06		7.14	8.05	7.24	6.44	6.01
SMC ISM[5]	10.91	7.73	6.63	8.03	7.27	6.59	5.81
LMC PN[6]	11.09	8.26	7.92	8.41	7.57	6.83	6.12
LMC ISM[5]	10.95	8.05	7.14	8.35	7.61	6.81	6.29
Galactic PN[7],[9]	11.05	8.39	8.14	8.69	8.10	6.91	6.38
Solar vic. ISM[5]	11.00	8.33	7.57	8.70	7.90	7.06	6.42
Sun[8]	11.00	8.60	8.00	8.93	8.09	7.21	6.56

[1]Howard, Henry & McCartney (1997) [2]Walsh *et al.* (1997) [3]Dopita & Meatheringham (1991a) [4]Dopita & Meatheringham (1991b) [5]Russell & Dopita (1992) [6]Dopita *et al.* (1997) [7]Kingsburgh & Barlow (1994) [8]Anders & Grevesse (1989) [9]Kwitter & Henry (1998).

These mean abundances should not be taken too seriously because there is a lot of scatter between objects, depending on the initial mass and age of the particular PNe. Nonetheless it is clear that the PNe have similar abundances in the $\alpha-$process elements, O, Ne, S and Ar as their local ISM, whereas He and N are strongly enhanced in the PNe shells (~ 0.13 dex for He, and ~ 0.6 dex for N). Carbon is also enhanced, but by a lesser amount.

Generally speaking, the carbon abundance obtained by observation of C II recombination lines such as C IIλ4267Å is always much higher than that inferred from the UV lines such as C III]λ1907,9Å lines (*i.e.* Kwitter & Henry, 1998). The reasons for this are not fully understood.

For the LMC PN Dopita *et al.* (1997) have been able to establish that the ratio of C to N dredge-up is a strong function of stellar mass and metallicity. The low mass, old stars with low initial metallicity appear to be much more efficient at dredging up C. On the other hand, the Type I PNe produced by young, massive stars, born from the chemically enriched ISM produce copious amounts of N by hot-bottom burning. This trend is seen clearly in Figure (10.4).

10.4.3 Nova Shells

Advanced Topic

Nova explosions result from the accretion of hydrogen-rich material derived from a binary companion onto the surface of a White Dwarf star. When enough gas has been accreted, thermonuclear hydrogen burning ignites once more. However, if the accretion rate has been low enough, the accreted material is strongly electron-degenerate. In these circumstances, the increase in temperature produced by the thermonuclear burning produces almost no
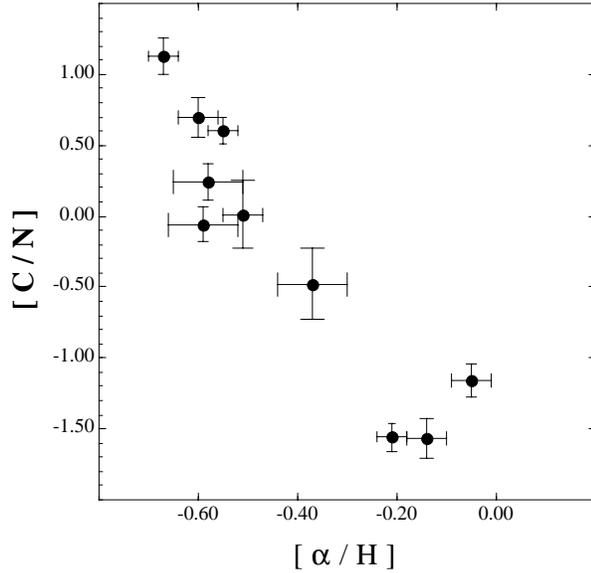
Fig. 10.4. The logarithmic C/N ratio (relative to the sun $= 0$) in the LMC planetary nebulae plotted against the α/H abundance. From the data of Dopita *et al.* (1997).

change in the pressure, which is dominated by the degeneracy pressure of the electrons. A *thermonuclear runaway* therefore occurs until the thermal pressure of the gas breaks the degeneracy, which means that the gas reaches temperatures of up to 10^8 K. The enormous and sudden release of thermonuclear energy causes the star to briefly shine at a luminosity of more than the Eddington value for the electron scattering opacity, and the radiation pressure drives a shell of gas out at velocities of hundreds to thousands of kilometers per second. During the ejection phase, the energy input from β^+-unstable nuclei such as ^{13}N, ^{14}O, ^{15}O, and ^{17}F is important in powering the luminosity and the outflow.

The nature of the explosion and the thermonuclear products produced depend on the mass of the white dwarf. If this mass is high ($> 1 M_\odot$, approximately), less mass can accumulate before thermonuclear runaway is initiated. As a consequence, the ejected mass is low (of order $10^{-5} M_\odot$), the velocity of ejection is high (several thousand km s^{-1}), and the variation in the emitted light is rapid (days). Such *fast novae* eject material rich in O, Ne and Mg. Low-mass white dwarfs produce slow novae, with light curves decaying over months, lower ejection velocities (a thousand km s^{-1}, or even less), and relatively large ejected masses (of order $10^{-4} M_\odot$). The ejected material is rich in the products of partial hydrogen burning, in particular, of N. Novae only produce about $10^{-3} M_\odot$ yr^{-1} of ejecta in the whole of the Galaxy, and are

therefore unimportant in the chemical evolution of the interstellar medium, with the exception of a few isotopic species of elements such as Al.

In the initial fireball phase, the nova shell is completely optically-thick and produces a (very hot) featureless continuum spectrum, however, as shell expands, the PdV work causes it to cool, and the peak of the emission shifts into the UV-visible regions of the spectrum, producing a maximum in the light curve. The more rapidly expanding outer portions of the fireball cool more quickly and start to recombine.

In this phase the hydrogen Balmer and the helium recombination lines are optically-thick and have a characteristic flat-topped appearance, with a flux limited by the black-body value. P-Cygni absorption features may be present on the blue wings of resonance lines, particularly in the UV, which are produced by the more rapidly expanding optically-thin material in the outer regions of the shell. During this phase, the very strong photon field at Lyβ can excite the O I fluorescence mechanism described in section (4.2.2). Again during this phase, dense clumps of recombined material may develop, possibly as a result of chemical inhomogeneities in the ejecta, since regions of high metallicity will cool and recombine more rapidly. Dust and molecules may also form at this point, which in some nova leads to the appearance of CO bands in the IR and a marked dimming at optical wavelengths. During this phase too the radiation diffusion timescale of the ejecta becomes short compared with the dynamical timescale of the expansion, so that from this point, the luminosity of the fireball cannot exceed the central source luminosity.

Eventually the ejecta become optically-thin in both lines and the continuum. This occurs first in lower density phase, and marks the transition to the *nebular phase* in which the photoionised shell is visible. This is marked by the appearance of PNe-like forbidden lines in the spectra. Initially, the density is so high, 10^7 - 10^8 cm^{-3}, that most species are strongly collisionally de-excited. Lines which are less affected, such as [N II]$\lambda5755$Å and [O III]$\lambda4363$Å, are particularly strong at this point. During the early nebular phase, the shell is re-ionised from within, so that the Hβ is constant. However when the ionisation front reaches the outer boundary of the ejecta, the shell becomes optically-thin, and the Balmer line flux falls according to what would be expected for an expanding shell of constant mass; $F(H\beta) \propto t^{-3}$. Photodestruction of the newly-formed dust may occur in this early nebular phase.

Eventually, the ionised shell becomes resolvable from earth, first in radio VLBI and later in the visible using HST or ground-based telescopes. Characteristically, the shell is elliptical and shows higher polar than equatorial expansion velocities. In this phase, the knowledge of the expansion velocity may be combined with the rate of angular expansion to derive a reliable distance estimate, so that absolute luminosities may be derived. Nebular photoionisation analysis shows this to be in the range 6×10^{37} erg s^{-1} up to 2×10^{38} ergs s^{-1}, comparable with the expected Eddington luminosity for electron-scattering opacity.

During the nebular phase and the final decline, the emission lines can be used to derive the nebular abundances in the same way as is done with PNe. Indeed, the abundances may also be estimated from the earlier optically-thick phase spectra, but these require the full methodology of radiative transfer in full non-LTE conditions in the intermediate density regime, which is difficult, and requires powerful supercomputers. The nebular optical spectra of the fast O, Ne, Mg novae are dominated by lines such as [Ne V]$\lambda\lambda$3300,3344Å, [Ne III]$\lambda\lambda$3868,3876Å, [O III]$\lambda\lambda$4363Å, [O III]$\lambda\lambda$4959,5007Å, [O II]$\lambda\lambda$7218,7328Å and recombination lines of He II and H. The slow novae have spectra much more like those of PNe, although the nitrogen lines are much stronger. The line profiles of both classes usually show the double-horn structure expected for an expanding shell, but there is often a lot of sub-structure caused by individual condensations.

The abundance analysis of nova shells has been led by Starrfield's group based at ASU. Since this group uses the same analysis technique, and combines the results of both UV and optical spectra in the analysis, their abundances are likely to be much better in a comparative sense, and are probably more accurate in an absolute sense as well. Some of these results, chosen to illustrate the variety of chemical enrichments which may occur, are summarised in Table (10.4.3). The references are [1] Schwarz *et al.* (1997), [2]Vanlandingham *et al.* (1997), [3] Vanlandingham *et al.* (1996), and [4]Vanlandingham *et al.* (1998). As can be seen by comparison with Table (10.4.2), some of the elements are enriched in their abundance relative to hydrogen by factors of a hundred or more!

The effective temperatures of the central stars are very high, 2 - 4 $\times 10^5$ K. The nebula is also very compact, typically 10^{15} to 10^{16} cm across. Thus, despite the high densities, the nebula is characterised by a high ionisation parameter, which helps to keep the electron gas hot.

Occasionally coronal lines are seen such as [Fe X]λ6374Å, or the [Si VII] lines in the IR. These may be indicative of shocks driven by a fast wind from the central star.

Element	He	C	N	O	Ne	Mg	Al	Si
Slow Nova:								
PW Vul (1984)[1]	11.00	9.39	9.90	9.65	8.07	7.58	-	-
Fast Novae:								
V 693 CorAus (1981)[2]	11.11	8.95	10.09	10.02	10.46	8.48	8.25	8.89
Nova Her (1991)[2,3]	11.11	9.22	9.15	8.35	9.67	-	-	-
Fast Nova in LMC:								
LMC 1990#1[4]	11.00	9.45	10.04	10.17	9.86	8.58	8.82	9.15

Over several years the surface of the white dwarf cools and fades, removing the power source for maintaining the photoionisation of the shell. The nebula therefore cools, and normally we would expect this to lead to rapid recombination. However, thanks to these extraordinary abundances, the cooling timescales in the plasma remain much shorter than the recombination timescales, at least until the electron temperature has fallen to the

point where the hyperfine line cooling starts to decrease, at a few hundred degrees K. Major cooling transitions at this time are the [O III]λ52,88μm, the [N III]λ57 μm and [N II]λ122,205μm lines.

In the optical, old nova shells display a curious spectrum dominated by hydrogen and helium recombination lines, as well as permitted lines of various abundant heavy elements which also arise from recombination (Williams *et al.* 1978; Williams, 1982). Weak [O II] and [N II] lines are visble, but these probably arise in a separate component with more normal abundances. The relative absence of the forbidden lines arises because the electron temperature is far too low to excite the visible forbidden lines. The Balmer continuum decreases very rapidly above the Balmer limit, which is caused by the $\exp\left[-h\nu/kT_e\right]$ term. This has been used by Williams to measure the electron temperature; of order 500 K for DQ Her and 800 K for CP Pup.

Because all the emission lines are the result of recombination, the ratio of the heavy element lines to the Balmer lines can be used to obtain the abundances, once the temperature is known. This needs low-temperature recombination coefficients, which were calculated and tabulated by Smits (1991). The calculation of absolute abundances would require a time-dependent photoionisation model so that ionisation correction factors could be properly calculated.

Notes on Chapter 10

- The non-LTE theoretical modelling of hot stellar atmospheres, including self-consistent treatments of mass-loss and spherical atmospheres has been developed over recent years by the Munich University group led by Kudritzki (see, for example, Sellmaier, F. *et al.* 1993, A&A, 273, 533 and references therein).

Exercises

Exercise 10.4.1. 1 A planetary nebula is excited by central star with a temperature of 10^5K, and it has an optical depth of 0.5 to ionising photons at the Lyman limit of hydrogen. Assuming that the central star has a Black-Body spectral distribution, calculate the mean energy per photoionisation of H, He, and He$^+$. Determine also the equivalent width of the Hβ line with respect to the stellar continuum at Hβ (assume that the nebular temperature is 10^4K).

11. Photoionising Shocks

"He saw, but blasted by excess of light,
Closed his eyes in endless night"
— Thomas Gray

Up to this point, we have only considered stars as the sources of photoionising photons. However, the cooling zones of shocks, and more especially, fast shocks, produce copious amounts of EUV or soft X-ray photons which escape both upstream and downstream to be absorbed in either a precursor photoionised zone, or in a narrow photoionised zone adjacent to the recombination region of the shock. In very fast shocks, the photon spectrum may be quite hard, so that the characteristics of the precursor photoionised region become quite difficult to distinguish from regions photoionised by power-law spectra, such are found close to active galactic nuclei (AGN). A necessary condition for the development of these photoionised zones is that the shock is fully radiative. For shocks with a velocities below 500 km s^{-1} in plasma of solar composition, the cooling timescale is given by equation (8.36):

$$\left[\frac{\tau_{cool}}{\mathrm{yr}}\right] \sim 200 \left[\frac{n}{\mathrm{cm}^{-3}}\right]^{-1} \left[\frac{v_s}{100 \mathrm{~km~s}^{-1}}\right]^{4.4} \tag{11.1}$$

However, at higher velocities than this, the mean cooling efficiency decreases as a result of magnetic support in the post-shock gas and as a result of the loss of the most efficient coolants in the highly ionised post-shock region; *see* Section (7.2), so that the cooling timescale becomes considerably longer, and less dependent on the shock velocity:

$$\left[\frac{\tau_{cool}}{\mathrm{Myr}}\right] \sim 1.9 \left[\frac{n}{\mathrm{cm}^{-3}}\right]^{-1} \left[\frac{v_s}{1000 \mathrm{~km~s}^{-1}}\right]^{2.9} \tag{11.2}$$

According to these equations, supernova remnants (SNR) evolving in the local interstellar medium ($n \sim 1$ cm^{-3}) will only become radiative when they have achieved a radius of 20-40 pc, and shock velocities of order 300 - 400 km s^{-1}. Many of the optical filaments seen in younger SNR are the results of shocks travelling into much denser clouds ($n \gtrsim 10$ cm^{-3}) with shock velocities of only 100 - 150 km s^{-1}, and the effects of photoionised shock precursors are therefore generally small. In order for the precursors to

be important, the supernova explosion must have occurred in an unusually dense environment, and the remnant would have to be quite young. Such conditions are encountered in starburst galaxies where the mean density of the interstellar gas can be as high as $n \sim 1000$ cm^{-3} and in which supernova remnants enter their radiative phase with diameters as small as 1 pc (Huang *et al.* 1994; Muxlow *et al.* 1994). For supernova remnants in unusually dense environments in our Galaxy and in the Magellanic Clouds, the photoionised precursors of fast shocks are occasionally seen (Shull, 1983; Dopita and Tuohy, 1984; Morse *et al.* 1996) and these provide a useful observational test of the theory.

By contrast, the jets of fast material shot out by active galactic nuclei (AGN) can persist over many millions of years, and the shocks that they generate may pass through several kiloparsecs of galactic interstellar matter before the jet escapes from the host galaxy into intergalactic space. In this case, shocks with velocities of 500 km s^{-1} or even 1000 km s^{-1} may become radiative, and their photoionising capability is considerable.

The idea that photoionising shocks might be important in the context of "broad-line" regions (BLR) of active galactic nuclei (AGN) dates back to Daltabuit & Cox (1972). Hybrid models of fast shocks and externally imposed photoionisation fields have been developed for application to the extended "narrow-line" emission regions (ENLR) of AGN in a series of papers by Viegas-Aldrovandi & Contini (1989, and references therein), but the detailed theory of steady-flow photoionising shocks was developed by Sutherland *et al.* (1993) and Dopita & Sutherland (1995,6) following an earlier attempt by Binette *et al.* (1985).

When the chemical abundance of the ejected material far exceeds that of the sun, the cooling timescales may be very much shorter, so that the shocks become promptly radiative. This applies in the oxygen-rich filaments ejected in Type II supernova explosions, which are seen when they are shock-heated by reverse shocks in their young supernova remnants. However, in this case the extreme abundances produce very peculiar photoionised regions, as we will see later.

11.1 Radiative Properties of Fast Shocks

11.1.1 Production of Photoionised Shock Precursors

According to the discussion of Section(8.2.3), a radiative fast shock will radiate essentially all of the kinetic energy flux across it, so that the total shock luminosity per unit area, given by equation (8.32) is:

$$\dot{E} = \frac{\rho_0 v_s^3}{2} \left(1 - \mathcal{M}^{-2} \right) \sim \frac{\mu m_H n_0 v_s^3}{2} \tag{11.3}$$

If the mean post-shock temperature in the cooling zone is T_1 then, using equation (8.24), the mean energy of the photons generated in the cooling zone is:

$$\langle h\nu \rangle \sim F(T_1)kT_1 = \frac{3}{16}f(v_s)\mu m_H v_s^2 \qquad (11.4)$$

where $F(T_1)$ and $f(v_s)$ are dimensionless functions of the shock temperature or shock velocity lying in the range 0 to 1. These functions are determined by detailed non-equilibrium cooling computations. If the cooling is dominated by collisional excitation line cooling in hydrogen and helium (and provided that the post-shock temperature is high enough that these species are fully ionised in the shock), $F(T_1)$ and $f(v_s)$ will decrease as T_1^{-1} (or v_s^{-2}), since the mean energy of the collisionally excited cooling lines remains constant with increasing shock velocity. On the other hand, in the limit that the cooling is dominated by electron Bremsstrahlung cooling, $F(T_1)$ and $f(v_s)$ will be constant, since the mean photon energy increases in proportion to the post-shock temperature. When the line cooling of heavy elements is included, the situation is intermediate between these two extremes, since the state of ionisation in the post-shock gas increases with temperature, as does the mean energy required to produce the collisionally excited lines. Since half of the photons generated in the cooling zone travel upstream, and the other half exit downstream, the total number of upstream photons produced per unit area of shock is, therefore:

$$S \sim \frac{\dot{E}}{\langle h\nu \rangle} \sim \frac{4}{3}n_0 f(v_s)^{-1}v_s \qquad (11.5)$$

This equation shows that the ionisation parameter in the pre-shock gas, $q = S/n_0$, increases with shock velocity at a rate which lies between $q \propto v_s$ and $q \propto v_s^3$. Detailed calculations by Dopita & Sutherland (1996) indicate that in the velocity range $200 < v_s < 500$ km s^{-1}, $q \propto v_s^{2.5}$ represents a good fit. This means that not only does the ionisation parameter increase faster than the shock velocity, but also that the ionising radiation becomes "harder" with increasing shock velocity.

Recalling (equation (9.6) *et seq.*) that q can be interpreted simply as the initial speed of advance of the ionisation front, it is clear that the condition for the production of a substantial precursor H II region is simply that $q > v_s$. The detailed calculations show that this is satisfied for shock velocities in excess of about 170 km s^{-1}, although the precise value depends on the abundance of the heavy elements in the shock. The dimensionless ionisation parameter in such precursor H II regions $\mathcal{U} = q/c \gtrsim 10^{-3}$, sufficient to ensure both that the precursor gas is fully ionised, and that the state of ionisation is quite high. The Strömgren column of such H II regions may be in excess of 10^{20} cm^{-2}, so that the precursor H II region is of order a hundred times more spatially extended than the cooling region of the shock.

When $q < v_s$, but $\mathcal{U} \gtrsim 10^{-5}$ there is still a photoionised precursor. However, in this case the photoionised region is trapped in a thin ionisation front

region located just in front of the shock front. The approaching atoms are ionised as they pass through this region, and all of the ionising photons are used up in producing new ionisations. The ionisation structure of such regions was first computed by Raymond (1976) and Shull & McKee (1979) and studied in detail by Cox & Raymond (1985). For a constant-velocity shock, pre-ionisation of hydrogen is essentially complete by about 100 - 120 km s^{-1}. Shocks which do not produce pre-ionisation are easy to distinguish, because the neutral hydrogen entering the high-temperature region behind the shock is collisionally excited several times before being ionised, producing a very strong hydrogenic two-photon continuum and a steep Balmer decrement (Dopita, Binette & Schwartz, 1982).

11.1.2 The Structure of Radiative Fast Shocks

Advanced Topic
    The computed structure of a fast shock, including its photoionised precursor is shown in Fig (11.1). This figure shows that photoionisation effects in the recombination zone of the shock are also important. Because fast shocks generate quite a hard photon spectrum, both of these photoionised zones show the effects of the penetrating soft X-ray photons, and their associated Auger ionisation which increases the fraction of C IV and CV ions, *c.f.* Sections (5.3.1) and (9.3).
    The ionisation parameter in the recombination region photoionised zone is lower than the pre-shock ionisation parameter by a factor equal to the compression factor in the shock. Generally speaking in the pre-shock gas, the thermal and magnetic field pressures will be in equipartition, so that the shock is supported by magnetic pressure in the recombination zone, and the compression factor is given by equation (8.34); $2^{1/2}\mathcal{M}_A$, where $\mathcal{M}_A$ is the Alfvén mach number of the shock, typically 15 - 50 for fast shocks with photoionised precursors. Thus, the ionisation parameter, which may be as large as $\mathcal{U} \sim 0.01$ in the pre-shock medium may only be of order $\mathcal{U} \sim 10^{-4}$ in the post-shock recombination zone. This means that the photoinised zone in the recombination region of the shock is of low excitation.
    The luminosity of the photoinised zones in the shock may represent an important fraction of the total shock luminosity in the IR, optical and far-UV regions of the spectrum. This is because an appreciable fraction of the shock luminosity in the EUV and soft X-ray regions of the spectrum is fluorescently converted into lines or continuum in the UV to the IR. Provided that the EUV photons are all absorbed in the shock structure, this fluorescence allows us to estimate the total mechanical energy flux through the shock. For example, the total luminosity in the [O III]λ5007Å line is computed to be 2% of the total mechanical energy flux. Other lines behave in a similar manner, although the fluorescent efficiency may be itself a function of the shock velocity, depending on how the overall spectrum evolves with shock velocity.
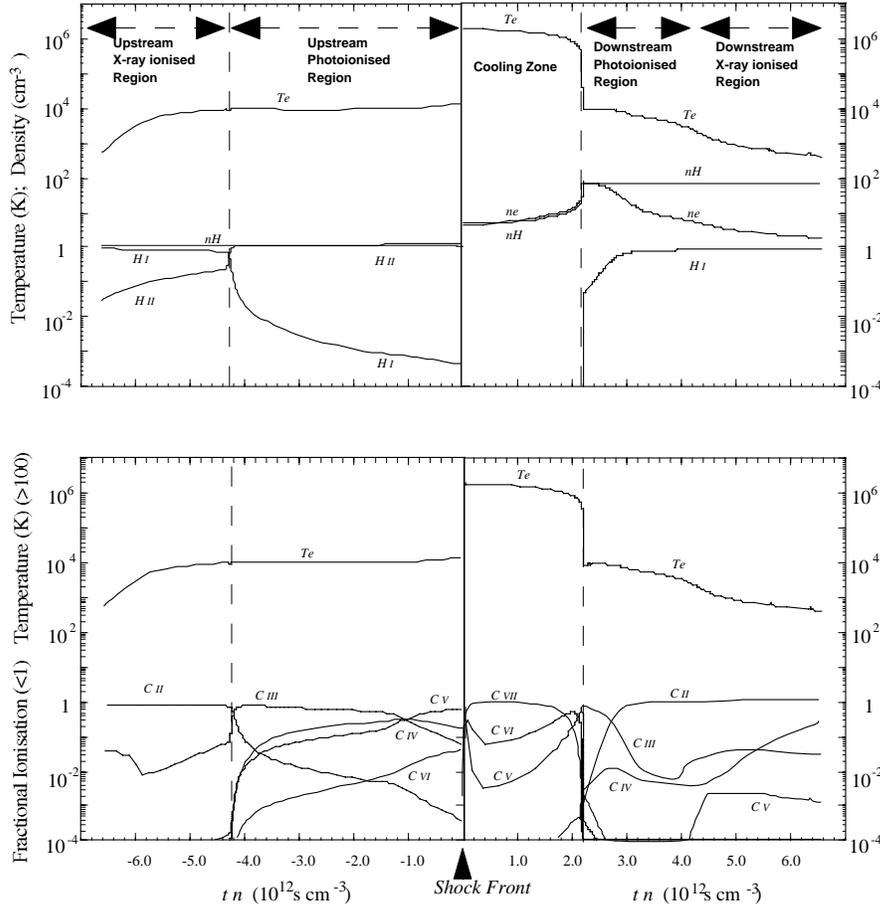
Fig. 11.1. The ionisation, temperature, and density structure of a 400 km s^{-1} radiative shock, after Dopita & Sutherland (1996). The x-axis is a time axis, scaled to be independent of density. This is better than a distance scale in making visible each of the shock zones, the approximate extent of which which are indicated at the top. The effects of Auger ionisation are evident in the C ionisation balance in the X-ray ionised regions of both the precursor and the recombination zone.

The nature of the emergent spectrum depends on a number of factors other than the Sonic and Alfvén Mach numbers of the shock. For example, the geometry of the situation producing the shock is an important parameter. If the shock arises in a geometrically thin disk of material, or if the shock results from the collision of two small clouds, there may not be enough material around to aborb the upstream ionising photons, so that the precursor H II region is truncated or incomplete. Because the ionisation parameter in the pre-shock gas is high, this loss not only lowers the overall luminosity, but also

tends to lower the excitation of the emergent spectrum. As limiting cases of the geometry, Dopita & Sutherland (1995,6) computed so-called "shock only" (cooling region and recombination region) spectra and complete "shock plus precursor" spectra.

Another important effect is the effect of the cooling instabilities. All fast shocks are thermally unstable, *see* Section (7.5), and this will tend to break up the recombination zone into a series of small cloudlets, as well as encouraging the development of a set of secondary shocks. In this case, it is the recombination region which fails to intercept the downstream photons, and the shock becomes more dominated by its photoionised precursor. In a cylindrical or spherical shock geometry downstream ionising photons may even cross the structure to be absorbed in the precursor of the shock on the other side. These effects have not yet been properly computed.

11.1.3 Narrow Line Regions in Active Galaxies

Advanced Topic

The overall properties of active galactic nuclear spectra are nicely reviewed by Peterson (1997). All classes of active galaxies (AGN) show some evidence of "narrow line region" (NLR) emission. Seyferts of Classes 1 and 2 (Khachikian and Weedman, 1974) display a rich spectrum of narrow lines, as do the narrow line radio galaxies (NLRGs). In these objects, the NLR is characterised by large [O III]λ5007/Hβ ratio ($\sim$7-20), high [O III] electron temperature (in those cases where it can be measured), the presence of He II and strong lines of [O I], [N II] and [S II]. In addition, coronal lines of highly ionised species such as [Ca V], [Fe VII], [Fe X] are often observed. Where such regions can be resolved, they are seen to come from a distinct and more restricted region close to the nucleus, rather than from the NLR as a whole.

As mentioned above, although the lines are termed "narrow", this is something of a misnomer, since the typical line velocity widths are in the range $200 < \text{FWHM} < 500$ km s^{-1}, and in a few objects the velocity dispersion even exceeds 1000 km s^{-1}. In Seyferts, where the NLR is spatially resolved, the emission is often, although not always, confined to a cone-shaped region emanating from the nucleus. In all of these senses, the ionisation cones of NGC 1068 represent the best-observed and brightest prototype of the high-excitation, high line width Seyfert 2 galaxies.

A second class of AGN, the low-ionization nuclear emission-line regions or LINERs, were first defined as a distinct class of AGN by Heckman (1980). The emission line ratio criteria given by Heckman define a distinct (if somewhat arbitrary) region of excitation space. According to his definition the [O II]$\lambda\lambda$3727, 3729 lines are stronger than [O III]λ5007; the [O I]λ6300/[O III]λ5007 ratio is less than about 0.33; and the [N II]λ6584/Hα ratio is larger than about 0.6. A survey of "normal" elliptical galaxies by Phillips *et al.* (1986) showed that low-level LINER activity is found in an appreciable fraction of these. The work of Ho, Filippenko, & Sargent (1995)

has shown that some low level LINER activity is even more ubiquitous than had been previously suspected in both elliptical and spiral galactic nuclei. However, some of these may not be true AGN. Indeed, all of the following types of objects have occasionally been classified as LINERs:

1. Emission in dense accretion disks associated with AGNs.
2. Shocks in accretion flows approaching AGN.
3. Shocks in outflows associated with AGN.
4. Emission from cooling flows.
5. Emission associated with old starburst nuclei.
6. Mixtures of emission due to circum-nuclear starbursts and Seyfert-like emission from AGN.

In the last twenty years, the hypothesis that all classes of narrow line regions (NLRs) are excited by a hard power-law spectrum of UV photons gained almost universal acceptance. In such a model, the Seyfert and QSO NLRs would be excited by a fairly flat UV power law, or else a truncated power law having an ionization parameter $\mathcal{U} \sim 10^{-2}$ (Koski 1978; Stasinska 1984; Veilleux & Osterbrock 1987; Osterbrock, Tran, & Veilleux 1992). On the other hand if LINERs are photoionized, the ionizing spectrum is similar to the NLRs, but the ionization parameter is low, $\mathcal{U} \sim 10^{-4}$ (Ferland & Netzer 1983; Ho, Filippenko, & Sargent 1993). The source of these photons is assumed to be the nucleus, but the exact means whereby the non-thermal UV spectrum is generated by the AGN remains obscure. Generally what is done is to simply join the observed soft X-ray to the observed EUV point, or to extrapolate the EUV slope with an exponential cutoff to match the soft X-ray. Neither of these procedures properly addresses the physical processes producing the emission, and there frequently remains a serious shortfall in the estimated number of ionizing photons based on an extrapolation of the observed UV spectrum.

Given that high-velocity outflows are a common feature of AGN, and that the NLR are dynamically rather active it is wise to consider the possibility that these regions might be excited by high-velocity photoionising shocks. This possibility was made more concrete by the models of Dopita & Sutherland (1995,6). They computed so-called "shock only" (cooling region and recombination region) spectra and complete "shock plus precursor" spectra and found that the former look very like LINER (low-ionisation emission region) objects associated with active galactic nuclei, while the latter look like Seyfert galaxy ENLR (extended narrow line region) spectra.

How may we distinguish emission regions photoionised by the AGN from emission from fast photoionising shocks? The most direct way is to look for differences in the emergent spectrum. In the optical, these differences are rather subtle, since the optical spectrum simply depends on the hardness of the ionising photon distribution and the value of the ionisation parameter. In the ultraviolet the lines become more sensitive to the existence

or otherwise of a hot, collisionally excited region such as can only be produced in the shock cooling zone. Best of all are temperature sensitive ratios, C III $\lambda\lambda977$Å/1907,9Åor the N III ratio, $\lambda\lambda991$Å/1750Åsuch as measured in NGC 1068 by Kriss $et\ al.$ (1992) using the Hopkins Ultraviolet Telescope. However, other UV or near-UV lines provide diagnostics, as demonstrated by Allen $et\ al.$ (1998). Figure (11.2) provides an example of such a diagnostic diagram.

Here, the theoretical "shock-only" and "shock + precursor" grids are given for shock velocities in the range 200 - 500 km s^{-1}, and for varying degrees of magnetic support, here represented by the pre-shock magnetic parameter $B/n_0^{1/2}$ (μG cm$^{-3/2}$). Photoionisation models are marked for two values of the slope of the (power-law) ionising spectrum, $\alpha = -1.0$ and $\alpha = -1.4$, two values of the ambient density, 100 and 1000 cm^{-3}, and a range of ionisation parameters $-3.5 < \log \mathcal{U} < 0.0$. On this, and on other plots, the M87 observations fall unambiguously into the "shock-only" region, but either pure photoionisation models or "shock + precursor" models can account for Seyfert II spectra.

In general, LINERs present an easier target in distinguishing between the excitation mechanisms, since the cooling zone of the shock, if present, directly affects the observed spectrum. A photoionised region producing a spectrum similar to a LINER must be characterised by low ionisation parameter. This in turn requires that the electron temperature is low, about 8000K or even less, and therefore UV lines have too high excitation energies to be excited to any large degree. On the other hand, a shock must always have a hot UV-emitting zone adjacent to the recombination zone. Thus, not only do spectra covering both optical and UV wavelengths allow one to distinguish between these excitation mechanisms (Dopita $et\ al.$ 1997), but also the UV spectra by themselves will reveal the continuum from any photoionising source, be it either an AGN or hot stars.

In addition to direct spectral diagnostics, there are a number of indirect methods of testing for shock excitation. These include:

- Correlations beween radio non-thermal and emission-line luminosity.
- Spatial correlations between radio non-thermal bubbles and emission-line regions.
- Correlations between excitation and dynamics, and
- Correlations between emission line fluxes and velocity dispersion.

As in most scientific debates, the truth of the matter is probably that both of the proposed mechanisms are important in exciting the extended emission line regions of AGN and high redshift radio galaxies.

11.2 The Oxygen-Rich Supernova Remnants

Advanced Topic
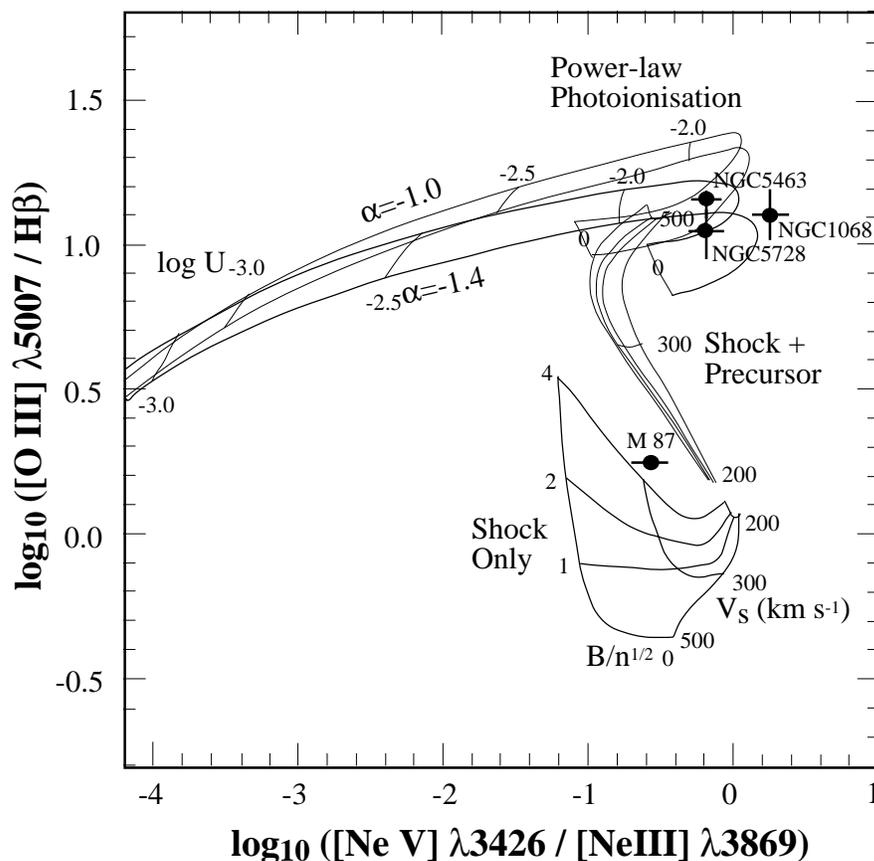
Fig. 11.2. An example of a diagnostic line ratio plot useful in distinguishing be-
tween photo-excitation from a central AGN, and excitation by fast shocks. The
three NGC objects represent the observations of Seyfert II galaxies, and position
of the LINER nucleus of M87 is also shown. (after Allen *et al.* , 1998

   The young supernova remnants (SNRs) provide an important key to test
the theories of nucleosynthesis in massive stars, since only in these objects
do we find material from the stellar interiors exposed to direct investigation.
This small class of SNRs shows fast moving (>1000 km s^{-1}) knots of material
emitting in lines of oxygen, neon, and other heavy elements, but apparently
containing no trace of hydrogen or helium. This is consistent with their origin
from within the helium-burnt layers of a massive ($> 12 M_\odot$) progenitor star.
In our Galaxy, Cas A is the prototype of this class (see Chevalier and Kirshner
1979). The other members galactic objects are G292+1.8 and Puppis A (Goss
et al. 1979; Winkler and Kirshner 1985). In addition, three such remnants are
known in the Magellanic Clouds (Lasker 1978; Dopita *et al.* 1981; Kirshner

et al. 1989), and two unresolved objects have been found in more distant galaxies M83 and NGC 4449 (Long *et al.* 1981, Blair *et al.* 1983).

For a long time, a fundamental barrier to a quantitative interpretation of the spectra from these objects was our inability to produce a plausible model for the excitation of these knots. It is clear that they were formed by a cooling instability in the supernova ejecta during the late fireball stage. Clear evidence for such instabilities has been found in the case of SN 1987A (DeKool, Li & McCray, 1998). The dense knots of oxygen-rich material so formed travel outward ballistically, and are excited at the time that they pass through the reverse shock which is propagated back into the low-density phase of the ejecta as this interacts with the surrounding ISM.

Logically, therefore, the knots are excited by slow shocks running into them which are driven by the increase in external pressure. However, the structure of these shocks is quite different from those which occur in material of normal composition. In particular, Itoh (1981a, b) and Dopita, Binette and Tuohy (1984) demonstrated that the very high metallicity causes the cooling time to remain shorter than the recombination time until the shocked plasma has cooled to a few hundred degrees K. As a consequence, the ionisation state becomes "frozen-in" the cooling plasma, and the emission lines of the lower ionisation species are supressed; contrary to observation.

Itoh (1981b) argued that the powerful EUV field generated in the cooling zone would escape upstream to produce an extensive warm photoionised precursor region. However, this idea runs into the fundamental difficulty that the extent of this region would have to be much greater than the observed size of the knots. Furthermore, if oxygen rich, the temperature of this region would high enough to excite optical forbidden lines only if it was composed of pure oxygen. The efficient cooling in the fine-structure lines of any other elements such as carbon produces a photoionisation equilibrium temperature of order 300 K, suppressing the cooling in the optical lines.

For the photoionised precursors in the oxygen-rich material, Dopita (1987) suggested a non-equilibrium solution. The buildup of the strong precursor radiation field occurs over the cooling timescale of the shocked plasma, allowing an ionisation front to be propagated into the cloud. For an $R-$ Type ionisation front driven by a strong photoionising field, the heating effect of the photon field dominates at the leading edge of the front, and the gas is strongly superheated, producing optical forbidden lines of a range of ionisation states.

This idea was fully developed by Sutherland and Dopita (1995b) who showed that the cloud shock by itself could drive the $R-$Type ionisation front, rather than the bow-shock emission which had been invoked by Dopita (1987). The buildup of the EUV emission from the cooling zone of the cloud shock is sufficiently rapid to allow an $R-$Type ionisation front to be detached from the shock front in roughly one tenth of the cooling timescale; about $3/n_0$ years, where n_0 is the pre-shock density in units of cm^{-3}. As this ionisation front initially starts to ionise the pre-shock plasma, the electron temperature

rises very quickly, since the energy input per photoionisation is large, but the absence of electrons ensures that the rate of cooling is relatively low. However, in such a heavy-element rich plasma, as the degree of ionisation increases the cooling rate increases at an even faster rate, which lowers the electron temperature again. Eventually the electron temperature falls back to its equilibrium level, about 300K, and only IR photons are produced. As a result, the optical emission is generated in a brief period during or immediately after the cloud shock becomes radiative, as the thin superheated zone close to the ionisation front sweeps through the cloud.

The ionisation front will pass through the cloudlet in a timescale comparable with the dynamical timescale, the time need for the cloud to fully enter the reverse shock. The dynamical timescale is itself a few times shorter than the cloud crushing and shredding timescale. For this reason, the radiative lifetime of an individual knot is short. For cloud densities $\sim 100 cm^{-3}$, it is comparable with the 25 year $e-$folding lifetimes observed for the Cas A knots by Kamper and van den Bergh (1976).

The combination of a (steady-flow) cloud shock of 100-200 km s^{-1} and its $R-$Type precursor gives a good description to the various ratios that can be formed from the [O I] $\lambda 6300$Å, the [O II] $\lambda 3727,9$Å and $\lambda 7316,24$Å, and the [O III] $\lambda 4363$Å and $\lambda 5007$Å lines. However, UV diagnostics from observations such as those by Blair *et al.* (1989) are required in order to obtain useful abundance data, especially for elements such as C, Mg and Si.

Notes on Chapter 11

- The field of active galactic nuclei is nicely reviewed in Peterson, B.M. 1997, *"An Introduction to Active Galactic Nuclei"*, CUP:Cambridge, ISBN 0-521-47911-8.

Exercises

Exercise 11.2.1. 1

12. Interstellar Dust

"The dust of exploded beliefs may make a fine sunset"
— Geoffrey Madan

It takes only an informed glance at the Milky Way on a dark and clear night to realise that interstellar dust exists. The luminous disk and bulge of faint Galactic stars is everywhere crossed by dark lanes, filaments and clouds of dense dusty gas made visible in projection. Because the dust grains are very small, typically less than a micron in diameter, their absorption characteristics are very wavelength dependent, and blue or ultraviolet photons are blocked and scattered much more easily than photons of longer wavelengths. The light from stars lying behind dusty interstellar clouds is therefore both attenuated and reddened in much the same way as smoke or industrial pollution causes the sun to appear red. The scattered light tend to be bluer than the illuminating star. In addition, anisotropic grains may be aligned, either by magnetic fields or some other agency, to produce an directional-dependent extinction which polarises the light in transmission or scattering.

Dust grains absorb the visible and UV light. Because they are cool, typically a few tens of degrees K, they re-emit this absorbed energy at far-infrared wavelengths. Thus, the Galactic distribution of dust can be mapped out using space-IR observations such as those by the IRAS and COBE satellites.

In external galaxies, the far-IR emission produced by grains is also very readily detected. Indeed some galaxies, identified by the IRAS satellite, which are undergoing violent and rapid bursts of dust-enshrouded star-formation may emit as much as 90% of their total energy output in the 10 - 300μmwaveband.

In dark clouds, the dust grains lock up an appreciable fraction of the elements having high condensation temperatures (*see* Figure 4.2). By adsorbing atoms, ions, and molecules onto their surfaces, they also provide ideal sites for facilitating complex interstellar chemical reactions which would never happen, or else happen only very slowly, in the gas phase.

The atomic bonds in the dust grains have characteristic frequencies for bending or stretching, depending on the atoms involved. These produce characteristic absorption or emission features in the mid-IR which have been successfully used to probe the nature and the chemical composition of the grains. One of the great successes of this technique has been the identification of an

abundant class of carbon-rich grains containing complex linked benzene rings. Nonetheless, many uncertainties remain in such analyses, and many questions on the details of the nature, size distribution and composition of the grains remain unanswered.

With this very short overview of the subject, let us now address the basic question of where the initial seeds of the grains we see in interstellar space were formed.

12.1 Stardust Formation

In principle, dust may be formed by direct condensation out of the gaseous phase whenever the kinetic temperature of the gas falls below the critical value which allows solids to form, and when a sufficient supersaturation in the gas phase exists for a timescale which is long enough. In practice, conditions for grain formation are realised in a variety of astronomical environments:

- In the wind around cool giant stars, either on the Giant Branch or on the Asymptotic Giant Branch (AGB) phase of evolution,
- In supermassive stars undergoing massive mass-loss episodes, such as in the star η Car,
- In nova envelopes, before the ejected shell becomes optically-thin to the escape of radiation, and
- In supernova ejecta when the diffusion timescale for the escape of stored radioactive energy becomes short compared with the expansion timescale of the fireball, allowing the ejecta to cool below 2000 K.

In all these cases, we have a gas which is slowly cooling from higher temperatures and in which the pressure is high enough to allow both nucleation and grain growth. However, in none of these cases are we likely to have a state of thermodynamic equilibrium in the gas which is forming the dust, and shock heating and cooling are often both important. Therefore, equilibrium chemical reaction networks for molecule and grain formation are inapplicable. We have to use instead a detailed time-dependent description of the chemical reactions, usually referred to as a *kinetic* model.

The gas-phase chemistry which occurs in these cases is vey complex, and the details are still uncertain (see, for example the review by Omont, 1991). However, our understanding of the broad features of grain formation appear to be secure. Two major paths to grain formation can be distinguished, depending on whether the outflow is carbon rich (in which case complex organic soots are created) or oxygen rich (in which case silicate grain types are dominant). These are discussed in more detail in the following sections.

Once a nucleation centre has formed, grains may grow by capturing further atoms or molecules from the gaseous phase. If the species being captured has a mass m, and is present at a density n, then the rate of increase in grain mass, m_g, is given by:

$$\frac{dm_g}{dt} = 4\pi a^2 \rho_g \frac{da}{dt} = \pi a^2 \left(\frac{8kT}{\pi m}\right)^{1/2} Smn \qquad (12.1)$$

where the thermal speed of the colliding species is $v = (8kT/\pi m)^{1/2}$, and the *sticking probability* per collision is S (*see* Draine 1978). This integrates to give the grain radius as a function of time:

$$a = a_0 + \frac{Sn}{\rho_g}\left(\frac{mkT}{2\pi}\right)^{1/2} \qquad (12.2)$$

The sticking probability is itself a function of temperature. According to the quantum mechanical theory worked out by Leitch-Devlin and Williams (1985), the approaching atom (*adatom*) is captured to the surface of the solid in a free-bound transition with the excitation of a lattice *phonon* (an excited vibrational state of the lattice). At low temperature, the gas atom has insufficient energy for phonon excitation, while at high temperatures, the energy of the phonon excited is insufficient to bind the adatom to the surface. The sticking probability therefore rises from zero at low temperature, reaches a maximum of around 0.2 - 0.8 when the mean thermal energy of the adatoms is about equal to the phonon energy, and then falls away at high temperatures. In practice, real grains present rough and irregular surfaces, which tends to enhance the sticking probability.

After being *adsorbed* to the surface, Hollenbach & Salpeter (1970) showed that the adatom migrates around the surface and progressively loses energy in a series of thermal hops or quantum-mechanical tunnelling events. The timescale for thermal hopping, τ_{hop}, is:

$$\tau_{hop} = \nu^{-1} \exp\left[\frac{\Delta E_{barr}}{kT}\right] \qquad (12.3)$$

where ν is the frequency for classical oscillation parallel to the surface, and ΔE_{barr} is the energy of the barrier that must be overcome to move on the surface. Thermal hopping will occur more readily the lighter the species, and for the lightest species, the quantum mechanical tunnelling rate can exceed the thermal hopping rate at low temperatures. For a rectangular barrier of width a, the tunnelling timescale, τ_{tun}, is given approximately by (Tielens & Allamandola, 1987):

$$\tau_{tun} = \nu^{-1} \exp\left[\left(\frac{2a}{\hbar}\right)(2m\Delta E_{barr})^{1/2}\right] \qquad (12.4)$$

This migration of species around the surface causes different adatom species to make encounters and so facilitates chemical interactions on the grain surfaces.

Because of the physics of the condensation process, and the interaction between the grains formed in the flow and the radiation field of the star,

there is a complex relationship between the nature of the grains, their size distribution and the terminal velocity of the outflow. Kozasa & Sogawa (1997) showed that the grain size increases as the mass-loss rate increases, since the size of the grain produced by condensation depends upon the gas density in the wind where a strong supersaturation exists in the gaseous phase and upon the period during which the condensation timescale is much shorter than the dynamical expansion timescale . On the other hand, radiation pressure acting upon the grains accelerates the stellar mass-loss flow (thereby arresting the condensation process). This has been seen observationally by Loup *et al.* (1993) and explained theoretically by Habing *et al.* (1994). The expansion velocities of the carbon rich objects are larger than those of the oxygen rich AGB stars, and radiation pressure induced expansion of the atmosphere may limit the size of the typical carbon-bearing grain to ~ 50Å, similar to that which is needed to explain the 2175Å bump in the interstellar extinction curve (see below).

The grain types that are likely to arise in oxygen-rich stellar atmospheres are discussed in the following section. In summary, the smaller grains may be of the aluminium oxide or calcium/aluminium silicate types, with the magnesuim silicates being confined to the larger grain population.

Both the oxygen rich and and carbon rich grain types will undergo further processing in the reaches of interstellar space through the various grain destruction processes, chemical photolysis by the UV radiation field, or mantle growth in dense interstellar clouds. We will discuss these processes in more detail later in this Chapter.

12.1.1 Grain Condensation in Oxygen-Rich Atmospheres

Advanced Topic

The pathways for grain formation in oxygen-rich atmospheres are very complex (see discussion by Tielens, 1990). Even the initial condensation process presents something of a problem. The first solid to condense directly from the gas phase is expected to be corundum, Al_2O_3. This occurs through the reaction:

$2Al + 3H_2O \rightarrow Al_2O_3 + 3H_2$; $T \lesssim 1760$ K.
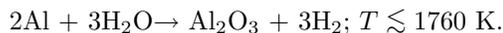
This is a rather improbable reaction thanks to the low gas-phase abundances (or partial pressures) of both of these species, and in particular, of the Al monomer. This reaction will have to occur close to the star, at about 3 stellar radii, since the effective surface temperature of the star is about 3000K. However, the actual condensation appears to occur in a region which is cooler than the critical condensation temperature of 1760 K by several hundred degrees. This lower temperature is necessary because clusters of several tens of monomers need first to form in order to get the stable solid seed about which further condensation can occur.

At slighly lower temperatures other rare-earth oxides such as CaO and TiO are expected to condense, and the TiO is converted to perovskite, $CaTiO_3$.

If the equilibrium condensation sequence were to be followed, which is appropriate to high-pressure environments with long dynamical timescales, gas-solid reactions are expected to produce the first silicate, gehlenite or melilite: $Ca_2Al_2SiO_7$. This reacts with gaseous Mg to produce diopside, $CaMgSi_2O_6$ and the Al released along with the remaining corundum is converted to spinel, $MgAl_2O_4$. Some of this reacts with the excess spinel in a solid-solid reaction to form the feldspar, anorthite, $CaAl_2Si_2O_8$ and the alkali metals are incorporated at lower temperatures still to form the alkali alumino-silicates: $NaAlSi_3O_8$.

In the dynamic conditions of red-giant atmospheres, this reaction sequence will "freeze out" at various stages, depending on the pressure, density, and dynamics of the wind. For example, a rapid drop in density will give rise to only high-temperature oxides, such as corundum, Al_2O_3 and perovskite, $CaTiO_3$. If the density remains high further out this will allow the formation of the magnesium silicates, forsterite Mg_2SiO_4 and enstatite, $MgSiO_3$ as thick mantles to the oxide grains. Iron may be incorporated into these mantles through gas-solid reactions with FeO to form the olivine, fayalite, Fe_2SiO_4, or else pyroxene, $FeSiO_3$. However, this will only occur if iron is still in the gaseous phase, and if the density is high enough in the region where these can survive. The formation of anorthite is unlikely, since the solid-solid reaction freezes out at very high temperature.

In conclusion, the theoretical composition of the smaller silicate grains ($a \sim 30$Å) in red-giant or AGB winds is likely to be $Ca_2Al_2SiO_7$ or $CaMgSi_2O_6$ while the larger grains ($a \sim 2000$Å or more) may consist of mantles of Mg_2SiO_4 and $MgSiO_3$ with some Fe_2SiO_4, or $FeSiO_3$ around cores made of Al_2O_3 or $Ca_2Al_2SiO_7$.

There is growing observational evidence that at least some part of this picture is correct. For example, amorphous silicates produce broad features at 9.7μmand 18μmwhich have been long known and observed in many objects. Recently, Waters *et al.* (1996) have used the short wavelength spectrometer (SWS) on the ISO satellite to identify a number of narrow features in supergiant, AGB and PNn spectra which are ascribable to crystalline silicates, in particular to pyroxenes (40.5μm), olivine (33.5μm, with additional features seen at 11.3μmand 23.5μm) or forsterite (32.8μm). These features are more prominent in the objects with cooler dust shells, suggesting that crystalline silicate formation is preferred in dense winds and high mass-loss rates, where the dust temperature is held above the glassification temperature for a long enough timescale to allow grain atoms to migrate into a crystalline matrix. In addition, the dense winds assist the incorporation of the iron into the silicate mantle.

12.1.2 Grain Condensation in Carbon-Rich Atmospheres

Advanced Topic

In a carbon-rich red giant or proto-planetary nebula atmosphere, the major carbon-bearing chemical that is initially formed is acetylene, C_2H_2, rather than the carbon monomer, thanks to the ready availability of hydrogen. Chemical nucleation from this precursor then leads to the formation of complex organic compounds. Léger & Puget (1984) were first to show that large planar molecules of polycyclic aromatic hydrodrocarbons (PAHs) were good candidates for the small grains produced in these C-rich outflows. PAH are linked series of aromatic benzene rings. These include members such as naphthalene $C_{10}H_8$ (two linked benzene rings), tetracene $C_{18}H_{12}$ (a linear string of four benzene rings), coronene $C_{24}H_{12}$ (a ring of six benzene elements around a central core), or ovalene $C_{32}H_{14}$ (eight benzene elements around two core elements). Such compounds have characteristic frequencies of absorption. They can, in principle, explain the characteristic emission features at $3.3\mu m$, $6.2\mu m$, $7.7\mu m$, $8.6\mu m$, and $11.3\mu m$ which are seen in all C-rich proto-planetary nebulae and planetary nebula envelopes (*i.e.* Aitken & Roche 1984, Cohen *et al.* 1986;1989). For example, the $3.3\mu m$ feature corresponds to C–H stretch modes in aromatic compounds, while the $6.2\mu m$ band is caused by the C=C stretch mode in aromatic rings (see Section (12.6.4) and Léger *et al.* 1991 for a more detailed discussion). However, no one type of PAH molecule provides a good fit to all of the IR features observed, suggesting that, although aromatic compounds are abundant in the carbon-rich grain species, their actual chemical makeup is rather complex.

The most likely chemical pathway which can convert acetylene to the first of the polycyclic aromatic hydrodrocarbons, naphthalene, is illustrated in Figure (12.1). The initial step is a three-body collision of acetylene forming propargyl radicals, which act to form the phenyl radical (benzene with a missing hydrogen atom). From this seed, alternating steps of hydrogen abstraction and acetylene addition produces napthalene, and can continue to produce yet more complex PAHs.

The emission features which can be ascribed to PAH-like molecules are observed to be ubiquitous in C-rich outflows. Indeed, these are the most abundant grain type in the interstellar medium, accounting for several percent of the carbon. This implies that the growth of PAHs must occur in a dynamical way, since otherwise the theoretical temperature range over which PAHs may form (1100-900K) is far too narrow to accommodate the growth which undoubtedly occurs.

The larger PAH molecules form flat plates that tend to agglomerate into a layered cluster structure. These finally become cross linked and clustered to form a quasi-spherical soot particle. For planar PAHs, the relationship between size, a, and number of C atoms they contain, N_C, is approximately $a \sim 0.9N_C^{1/2}\text{Å}$, while the larger clusters tend to the spherical relation, $a \sim 1.3N_C^{1/3}\text{Å}$. When the PAHs have grown to more than about 500 C atoms
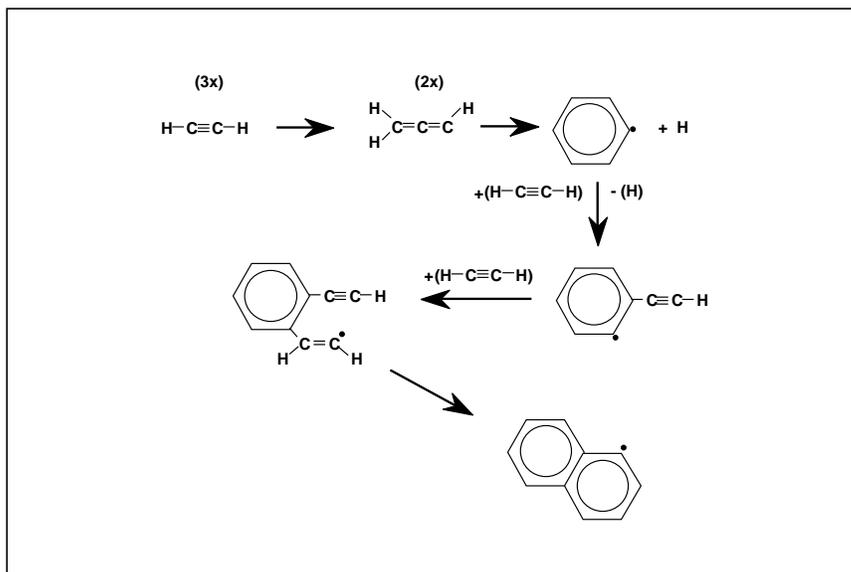
Fig. 12.1. The formation of PAHs in C-rich stellar outflows occurs first by the formation of the phenyl radical, from acetylene, followed by alternating steps of abstraction of H and the addition of acetylene (after Frenklach & Feigelson, 1989).

($\langle a \rangle \sim 10\text{-}20\text{Å}$), the force of radiation pressure is expected to become significant. Eventually this becomes large enough to drive an outflow, bringing the growth phase to an end (Cherneff *et al.* 1992).

12.2 Optical Properties of Dust

The presence of dust grains in the interstellar gas is inferred by the effect they have on the transmission of electromagnetic waves through interstellar space. Grains absorb, scatter, polarise and re-emit radiation, are heated and photoelectrically charged by energetic photons, and are accelerated and (possibly) aligned by radiation pressure forces. Each of these processes is strongly dependent on the nature of the grain material, and upon the wavelength of the absorption or re-emission. In the following sections, we will see how each of these processes operates.

12.2.1 Extinction : Definitions

As a beam of light from a distant star passes through a column of dusty gas, light is removed from the beam. The optical depth for interstellar extinction

at any wavelength is defined by the fraction of the incoming flux that remains in the beam:

$$I(\lambda, \tau) = I_0(\lambda) \exp\left[-\tau_{ext}\right] \tag{12.5}$$

Light has been removed from the beam in two ways, either by direct absorption onto the dust grain surface, or by scattering out of the beam. The extinction optical depth is therefore the sum of these two components:

$$\tau_{ext} = \tau_{abs} + \tau_{sca} \tag{12.6}$$

To proceed further, we now have to introduce a set of definitions which are used in the definition of the wavelength-dependent absorption properties of dust. Regrettably, astronomers do not use physical units and still insist on measuring the brightness of stars in terms of *magnitudes*, units historically related to the sensitivity of the human eye to detect differences in brightness. Thus a Ptolemy (or indeed, a current-day observer) would just be able to distinguish between two stars differing by 1/10 of a magnitude, or roughly 10% difference in terms of their flux. The magnitude system is therefore a logarithmic scale, and one magnitude change has now been (precisely) defined as a difference of $10^{0.4}$ in flux. The zero of the magnitude scale at all wavelengths is defined to be the brightness of the star Vega (α Lyr), and fainter stars have positive *apparent magnitudes*. Because the intrinsic spectral energy distribution of this reference star is not flat in energy units, the magnitude system also does not measure absolute energy – another factor to be regretted!

The *absolute magnitude, $M(\lambda)$*, of a star is defined as the magnitude that the star would have if placed at a distance of ten parsecs from the earth. Assume that the total (wavelength- dependent) absorption due to dust is $A(\lambda)$ magnitudes. From the above definition of the magnitude scale, this is related to the extinction optical depth by $A(\lambda) = 0.4^{-1} \log_{10}[e]\ \tau_{ext}$ or $A(\lambda) = 1.086\ \tau_{ext}$. The apparent magnitude, $m(\lambda)$ of a star with distance D parsecs from the earth is then:

$$m(\lambda) = M(\lambda) + A(\lambda) + 5\log D - 5 \tag{12.7}$$

By observing at only one wavelength, we clearly cannot sort out how much of the apparent magnitude is the result of the absolute magnitude of the star, how much is the result of distance, and how much is the result of dust absorption. However we can use the spectroscopic signature (the depth, ratios and widths of the absorption lines) to determine the stellar spectral type, which gives both its absolute magnitude and its intrinsic colour. By observing at two different wavelengths, we can then use the fact that the dust extinction is highly wavelength dependent to isolate the distance and the extinction-dependent terms in equation (12.7). We then have a measure of the *colour excess*:
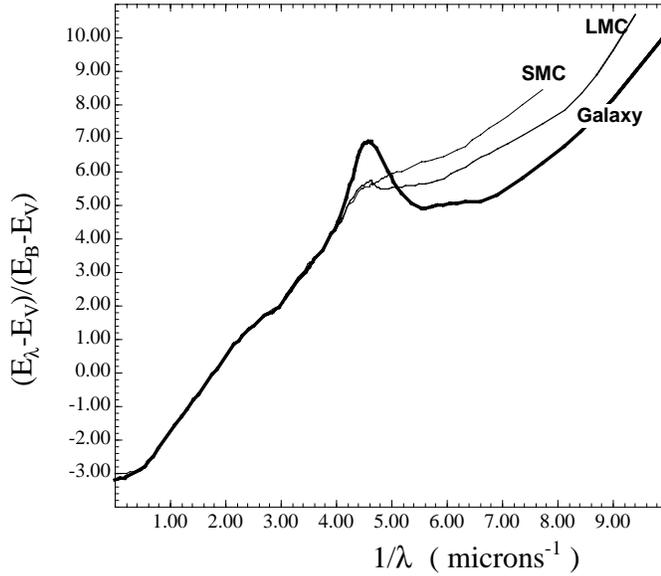
Fig. 12.2. The mean dust extinction curves for the Galaxy, the LMC and the SMC, drawn from the references cited in the bibilography. Within any galaxy, the intrinsic variability of the extinction curve is at least as great as the differences between these curves.

$$E_{\lambda_1} - E_{\lambda_2} = A\left(\lambda_1\right) - A\left(\lambda_2\right) \tag{12.8}$$

Normally for stars the visual, V, and the blue, B, magnitudes on the Cousins photometric system are known. These correspond to effective wavelengths of 5470Å and 4340Å, respectively. We therefore normalise the measured colour excess with respect to these:

$$\frac{E_{\lambda-V}}{E_{B-V}} = \frac{A\left(\lambda_1\right) - A\left(V\right)}{A\left(B\right) - A\left(V\right)} \tag{12.9}$$

The ratio of the *total to selective absorption*, R, can then be defined by:

$$R = \frac{A\left(V\right)}{E_{B-V}} = -\left[\frac{E_{\lambda-V}}{E_{B-V}}\right]_{\lambda\to\infty} \tag{12.10}$$

In the optical region of the spectrum, the extinction varies as λ^{-1}, approximately, so it is common practice to plot the extinction function defined by equation (12.9) in terms of inverse λ units (μm^{-1}). This is called the *reddening law* (appropriate to that particular line of sight). Typically, both the reddening law and the derived value of R are dependent on the environmental conditions along that line of sight, since they depend upon the chemical mix, physical composition and size distribution of the grains. Typically, $R \sim 3.1$

in the Galaxy, but it can vary between 2.7 and about 4.5. In general, regions with flatter extinction laws in the UV are characterised by larger R values.

The variability between the mean extinction laws for the Galaxy, the Large Magellanic Cloud and the Small Magellanic cloud is shown in Figure (11.2). However, within each galaxy, wide variations in the local extinction are observed (*see* Prévot *et al.* 1984, Rodrigues *et al.* 1997, SMC; Fitzpatrick, 1986 and Schwering & Israel 1991, LMC; and Fitzpatrick & Massa, 1990, the Galaxy). The sense of main variation in the extinction law is that the UV extinction law becomes steeper and the broad 2175Å absorption peak becomes weaker as we go to regions of lower density and of lower metallicity. Even the galaxy M31 appears to conform to this behaviour (Bianchi *et al.* 1996). A simple but accurate parametric fitting of the UV extinction law was developed by Fitzpatrick & Massa (1988, 1990):

$$\frac{E_{\lambda - V}}{E_{B - V}} = c_1 + c_2 x + c_3 D\left(x : \gamma, x_0\right) + c_4 G(x) : x \geq 3.3 \mu m^{-1} \quad (12.11)$$

$$x \quad = \lambda^{-1} : D\left(x : \gamma, x_0\right) = \frac{x^2}{\left[(x^2 - x_0^2) + x^2 \gamma^2\right]}$$

$$G(x) \quad = 0.5392\left(x - 5.9\right)^2 + 0.05644\left(x - 5.9\right)^3 : x \geq 5.9 \mu m^{-1}$$

$$= 0 : x < 5.9 \mu m^{-1}$$

The term $D\left(x : \gamma, x_0\right)$ may be recognised as a Drude resonance profile, closely related to the Lorentzian resonance profile; *c.f.* Equation (4.1), but applicable when the width of the profile becomes an important fraction of the resonant frequency. Indeed, this is what the 2175Å absorption feature appears to be: a surface charge resonance on very small carbon grains with radii $\sim 200 \mathring{A}$. The c_1 and the $c_2 x$ factors describe the absorption produced by the larger grain population, while the $c_4 G(x)$ factor results from a population of very small dielectric grains. The difference between the SMC, LMC and Galactic curves may therefore be described in terms of a relative absence of the type of carbon grains that are carriers of the 2175Å feature, along with an underabundance of larger grain types in the LMC and (even more so) in the SMC. It is not yet clear how much of this variability in the extinction curve is due to differences in chemical makeup, how much due to differences in the grain condensation processes, and how much is due to differences in the environment (UV radiation fields, grain destruction rates *etc.*). Certainly, the LMC is deficient in heavy elements by about 0.3 dex with respect to the local interstellar medium, and the SMC is even more deficient; by about 0.6 dex. As a result, the dust extinction per unit mass of gas is lower. This is shown by the relative values of the (somewhat inaccurately named) *gas-to-dust ratio*;

$$\frac{N(H)}{E_{B - V}} \sim 5.9 \text{x} 10^{21} \text{cm}^{-2} \text{ mag}^{-1} : \text{Galaxy}$$

$\sim 2.4 \mathrm{x} 10^{22} \mathrm{cm}^{-2} \mathrm{~mag}^{-1}$: LMC

$\sim (0.7 - 4) \mathrm{x} 10^{22} \mathrm{cm}^{-2} \mathrm{~mag}^{-1}$: SMC

where the values for each galaxy have been estimated by Bohlin *et al.* 1978, Fitzpatrick, 1986 and Rodrigues *et al.* 1997).

The value of the reddening constant, R, and the various factors in equation (12.11) tend to be correlated, to greater or lesser extent. For example, regions with flat UV extinction (low c_4) are seen to have larger values of R, and such regions occur in lines of sight passing through denser clouds. Regions with steep UV extinction laws tend to have wider 2175Å features (Fitzpatrick & Massa, 1986).

In the measurement of nebular emission-line spectra, it is not convenient to normalise the spectrum to the V band, and it makes a good deal more sense to reference the spectrum to the intensity of a Balmer line when de-reddening the observed spectra. Therefore, for this purpose we use an extinction law referenced to the *logarithmic reddening constant* defined at the wavelength of Hβ, c. If $F(\lambda, \tau)$ is the flux of a line of wavelength λ, and $F_0(\lambda)$ is its reddening-corrected flux, referenced to Hβ, then;

$$F(\lambda) = F_0(\lambda) \, 10^{-cf(\lambda)} \tag{12.12}$$

With a standard Galactic reddening law, $f(V) = 0.8643$, so that $A(V) = 0.8643 \times 10^{0.4} c = 2.161 c$.

12.2.2 Optical Efficiency Factors

The extinction and scattering due to grains, and its wavelength dependence provide a fundamental diagnostic tool to examine the grain composition and size distribution. Suppose that light of a particular *wavenumber*, $k = 2\pi/\lambda$, and of intensity I_0 is scattered by a dust particle. The intensity of the light, I, at a distance r from the scattering particle, and at angle θ with respect to the incident direction, and at azimuthal angle ϕ is then:

$$I(k) = I_0(k) \frac{f(\theta, \phi)}{k^2 r^2} \tag{12.13}$$

where $f(\theta, \phi)/k^2$ is the angular scattering distribution. The total scattering cross - section of the particle is then the integral of this distribution over all solid angles:

$$C_{sca} = \frac{1}{k^2} \int f(\theta, \phi) \, d\omega \tag{12.14}$$

It is convenient to compare this cross - section with the geometrical cross - section of the scattering particle; πa^2, where a is the radius, assuming a spherical particle (or, if elliptical, πab). We then have the *scattering efficiency* of the particle defined as:

$$Q_{sca} = \frac{C_{sca}}{\pi a^2} \tag{12.15}$$

In a similar way we can define the absorption efficiency Q_{abs} from which it follows that the *extinction efficiency* is:

$$Q_{ext} = Q_{abs} + Q_{sca} \tag{12.16}$$

This is simply related to the optical depth for extinction defined in the previous section:

$$\tau_{ext} = \int_{a_{\min}}^{a_{\max}} Q_{ext} n(a) \pi a^2 da \tag{12.17}$$

The *albedo* of a particle is defined as $= C_{sca}/C_{ext} = Q_{sca}/Q_{ext}$.
The *mean scattering angle*, or *asymmetry factor*

$$\langle \cos\theta \rangle = \frac{1}{C_{sca} k^2} \int f(\theta, \phi) \cos\theta d\omega \tag{12.18}$$

determines the momentum transfer imparted to the grain (dp_g) by the incident radiation field of intensity $I(k)$:

$$\frac{dp_g}{dt} = \frac{1}{c} \int_k I(k) \left[C_{abs} + (1 - \langle \cos\theta \rangle) C_{sca} \right] dk \tag{12.19}$$

The scattering efficiencies are calculated through the Mie theory of small-particle scattering. This is given in detail in van der Hulst (1957). Here we will only consider two useful limits; when the particle is much smaller than the wavelength of the incident light, and when it is large in comparison with this limit.

The larger particles have scattering cross - sections which are exactly twice the geometrical cross sections; $Q_{sca} = 2$. This factor arises because of the effect of the Fraunhofer diffraction by the grain, which adds to the geometrical cross - section. Fraunhofer diffraction applies in the far field limit: $r/a \gg a/\lambda$, where the geometrical shadow cast by the grain has been overwhelmed by diffraction effects.

For $a \ll \lambda$, we can regard the particle as sitting in a time-dependent but, from the particle's viewpoint, spatially uniform electric field, E, of the incident plane-polarised radiation. This field then induces a dipole moment in the particle, $p = \alpha E$, where α is the polarisability. This dipole moment then varies in step with the electric field of the passing electromagnetic wave, and radiates according to classical theory. For a uniform sphere, the polarisability is given by

$$\alpha = \left(\frac{m^2 - 1}{m^2 + 2} \right) a^3 : m = n - i\kappa \tag{12.20}$$

where m is the complex index of refraction of the grain material at that particular wavelength. In this case the scattering efficiency is:

$$
\begin{aligned}
Q_{sca} &= \frac{8\pi}{3} k^4 \left| \alpha \right|^2 \frac{2}{\pi a^2} \\
&= \frac{8\pi}{3} \left(\frac{2\pi a}{\lambda} \right)^4 \left| \frac{m^2 - 1}{m^2 + 2} \right|^2
\end{aligned}
\tag{12.21}
$$

If m is real (zero absorption), the extinction efficiency reduces to the Rayleigh Scattering (small particle) limit. If m is complex, we have an absorption efficiency:

$$
Q_{abs} = -4 \left(\frac{2\pi a}{\lambda} \right) \mathrm{Im} \left[\frac{m^2 - 1}{m^2 + 2} \right]
\tag{12.22}
$$

which, because it varies as λ^{-1} rather than as λ^{-4} for the scattering, will dominate the extinction at sufficiently long wavelengths.

For a single grain, therefore, the scattering efficiency rises from a small value at long wavelengths, reaches a maximum in the vicinity of $\lambda \sim 2\pi a$, and, as the wavelength of the incident light continues to increase, oscillates with decreasing amplitude about $Q_{sca} \sim 2$.

At sufficiently short wavelengths, (X-ray wavelengths) the radiation interacts with the individual atoms in the grain, rather than with the grain material in bulk, so that the opacity of the grains reduces to the photoelectric X-ray opacity; *see* Section (5.3.2). This means that X-ray absorption measurements of un-ionised dark dust-containing clouds can be used to establish the column density of gas in those clouds, once we think we know the relative abundance of the heavy elements. Such X-ray shadowing measurements can be used, for example, to indirectly study the total (atomic plus molecular) hydrogen fraction of such clouds.

12.2.3 Polarisation by Dust

If the light incident on a grain, $I(\lambda)$ is originally unpolarised, then it can be resolved into two incoherent beams polarised at right angles, and each with the same intensity $I_1(\lambda) = I_2(\lambda) = I(\lambda)/2$. If the grains are prefentially aligned along one of these directions, then the two beams will undergo different attenuations. Assume that the attenuation of the second polarised beam is larger than the first. Then, analysis of the transmitted light with a linear polariser will then give a maximum intensity, $I_{\max}$ and a minimum intensity $I_{\min}$ given by:

$$
\begin{aligned}
I_{\max} &= I_1(\lambda) \exp\left[-\tau_1(\lambda) \right] \\
I_{\min} &= I_2(\lambda) \exp\left[-\tau_2(\lambda) \right]
\end{aligned}
\tag{12.23}
$$

The *degree of polarisation* is defined by $p = (I_{\max} - I_{\min})/(I_{\max} + I_{\min})$, so it follows that the measured degree of polarisation is:

$$
\begin{aligned}
p(\lambda) &= \frac{\exp\left[-\tau_1(\lambda)\right] - \exp\left[-\tau_2(\lambda)\right]}{\exp\left[-\tau_1(\lambda)\right] + \exp\left[-\tau_2(\lambda)\right]} \\
&= \frac{1 - \exp\left[-\Delta\tau(\lambda)\right]}{1 + \exp\left[-\Delta\tau(\lambda)\right]}
\end{aligned}
\tag{12.24}
$$

where $\Delta\tau(\lambda)$ is the difference in the optical depths of dust as seen by the two beams. This difference is simply related to the efficiency factors for extinction in the electric field plane and the magnetic field plane, $\mathbf{E}$ and $\mathbf{B}$, respectively:

$$
\frac{\Delta\tau(\lambda)}{\tau(\lambda)} = \frac{|Q_{ext}(\mathbf{E}, \lambda) - Q_{ext}(\mathbf{B}, \lambda)|}{Q_{ext}(\mathbf{E}, \lambda) + Q_{ext}(\mathbf{B}, \lambda)} = \frac{p(\lambda)}{2.172A(\lambda)}
\tag{12.25}
$$

Equation (12.25) emphasises the very close relationship between the extinction produced by a given grain, and the polarisation which results. It is clear for example that, for an elongated grain, the polarisation reaches a maximum when the difference betwen the extinction in the long axis direction and the short axis direction also reaches a maximum, which will occur near the middle of the rapidly-rising section of the extinction curve. This behaviour is shown schematically in Figure (12.3).

When the integral is taken over the grain-size distribution, the wiggles caused by interference effect disappear, and the polarisation curve becomes a smooth function of wavelength. An empirical fit to the wavelength dependence of interstellar polarisation is provided by Serkowski relationship, which in the modified form provided by Whittet *et al.* (1992) is:

$$
\begin{aligned}
p(\lambda) \quad &= p_{\max}\exp\left[-K(\lambda_{\max})\ln^2\left(\frac{\lambda_{\max}}{\lambda}\right)\right] \\
K(\lambda_{\max}) &= 1.66\lambda_{\max}
\end{aligned}
\tag{12.26}
$$

where $\lambda_{\max}$ is the wavelength μm, where the polarisation reaches its maximum value, $p_{\max}$. Because the shape of the polarisation curve and the wavelength of the maximum is determined by the size distribution of the grains, (in particular, by the largest grain population present) it is therefore not surprising that the wavelength of maximum is related to ratio of the total to selective extinction, R; *see* equation (12.10). Whittet & van Breda (1978) find $R = 5.6 \pm 0.3\lambda_{\max}$.

In the ultraviolet, the absorption of the graphite grains producing the 2175Å resonance absorption feature must be taken into account. This has been done using the discrete dipole array (DDA) method of calculation for electromagnetic scattering developed by Draine (1988). This produces an additional bump in the polarisation around the 2175Å feature. The DDA
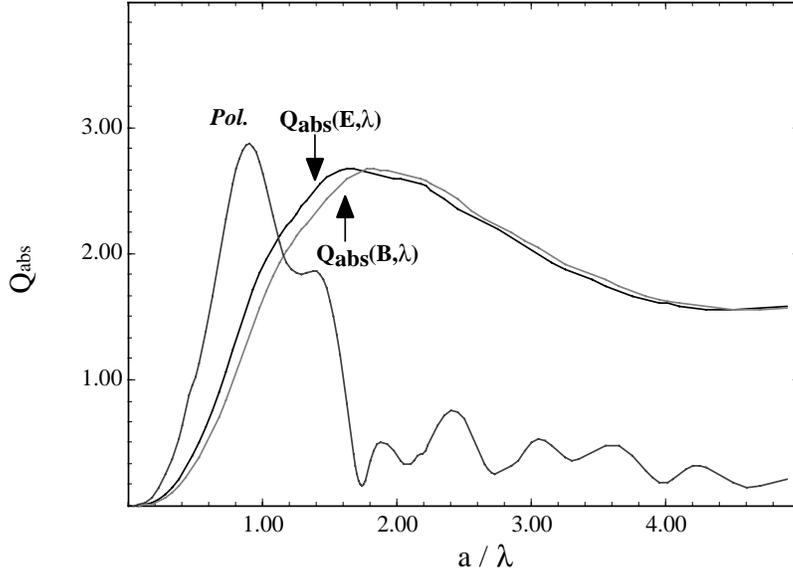
Fig. 12.3. A typical absorption efficiency curve for an aligned interstellar grain. The polarisation resulting from the difference of the **E** and the **B** absorption efficiencies is scaled arbritarily for clarity. The wiggles are the result of interference effects, and are smoothed out when there is a distribution of grain sizes present.

method is a much more powerful technique than the traditional Mie scattering theory, since it can readily deal with grains of arbitrary shape, or complex fractal structure which are much more like the real grains found in interstellar space.

12.3 Grain Photoheating

Interstellar grains are heated by photons or by atomic collisions and re-radiate this heat at infrared wavelengths. Let us first consider only the heating by the ambient radiation field. Suppose that this interstellar radiation field has a uniform energy density U (erg cm^{-3}) then the flux of energy, dF, intercepted by an elementary area of the grain ds is:

$$dF = \frac{c}{4\pi} U \, ds \int_{0}^{\pi/2} 2\pi \sin\theta \cos\theta d\theta = \frac{c}{4} U \, ds \qquad (12.27)$$

where θ is the angle of the local normal to the grain. Note that any element of the surface can only see half of the sky. Thus the net energy transport to the grain, dE_g/dt, is the integral of this over the surface of the grain, multiplied

by the fraction of the radiation which is absorbed, which for a spherical grain of radius a is:

$$\frac{dE_g}{dt} = \pi c a^2 \langle Q_{abs} \rangle U \qquad (12.28)$$

where $\langle Q_{abs} \rangle$ is the mean (frequency weighted) absorption efficiency.

The grains are heated to a temperature at which they are able to re-radiate this energy, so in thermal balance:

$$\pi c a^2 \langle Q_{abs} \rangle U = 4\pi a^2 \int\limits_0^\infty Q_{em}(a, \lambda) B(\lambda, T_g) \, d\lambda \qquad (12.29)$$

where $B(\lambda, T_g)$ is the Black-body function appropriate to the bulk grain temperature T_g, and $Q_{em}(a, \lambda)$ is the emission efficiency of the grain, which we can take equal to the absorption efficiency. Since most of the radiation is emitted in the IR where $n \sim 1$ and $\kappa \ll 1$, then equation (12.22) reduces to:

$$Q_{em}(a, \lambda) \sim \frac{16\pi a \kappa}{3\lambda} \qquad (12.30)$$

A rough idea of the typical grain temperature can be got by evaluating the average $\langle Q_{em} \rangle$ at the peak of the Black body distribution ($\langle Q_{em} \rangle \sim 8\pi a \kappa T_g / 3$), and taking it out of the integral. We also take $\langle Q_{abs} \rangle \sim 1$, since the local stellar radiation field is essentially diffuse starlight, and has a colour temperature of about 10000K, so that we have:

$$T_g \sim 0.4 \left(\frac{cU}{a\kappa\sigma} \right)^{1/5} \qquad (12.31)$$

where σ is the Stefan's constant. The local interstellar radiation field has $cU \sim 2 \times 10^{-2}$ ergs cm^{-2} s^{-1} (Mathis, Mezger & Panagia 1983), and therefore a grain with $a \sim 2 \times 10^{-5}$ cm will reach a temperature of about 20K. Note that this temperature is only very weakly dependent on grain size. The grain temperature depends on both the nature of the grain and upon the local radiation field. In ionised plasmas in starburst regions the dominant contribution to the local radiation field is the nebular Lyα, or, nearer to the exciting stars, the stellar radiation field due to hot stars. In such high radiation field densities the mean grain temperature may approach 100K.

12.3.1 Quantum Heating

In very small grains, the quantum heating due to the absorption of individual photons may cause the grain temperature to fluctuate, with large excursions above the mean temperature predicted by an equation such as (12.31). This occurs when the photon energy becomes comparable with the mean internal lattice energy of the grain carried in its vibrational modes, as was first shown

by Duley (1973) and by Purcell (1976). More recent calculations by Draine
& Anderson (1985), Desert *et al.* (1986) and Guhathakurta & Draine (1989)
have quantified both the size of the fluctuations and their effect on the IR
emission from dust. If the lattice contains N atoms, then the absorption of
each photon can excite $(3N - 6)$ or $(1 - 2/N)$ vibrational modes. Following
an absorption of a photon of energy $h\nu$ therefore, the grain is heated to a
maximum temperature T_{max}, and the internal energy of the grain, U, is:

$$U\left(T_{\mathrm{max}}\right) = (1 - 2/N)V \int_{0}^{T_{\mathrm{max}}} C_V\left(T\right) dT \sim h\nu \qquad (12.32)$$

where $C_V\left(T\right)$ is the heat capacity at constant volume of the bulk material.
The grain will cool on a characteristic timescale:

$$\tau_{cool}(T_g) = \frac{U\left(T_g\right)}{4\pi a^2 Q_{em}\left(T_g\right)\sigma T_g^4} \qquad (12.33)$$

Because of the T_g^4 term and the decreasing emission efficiency of the grain as
a function of temperature, the cooling timescale is initially short, but rapidly
increases as the grain cools. The temperature fluctuations have therefore a
very "spiky" characteristic as a function of time. Hot small grains contribute
very strongly to the total IR emission due to dust in the $\sim$5 - 25 μmwavelength
region.

   If the photon energy is large enough, T_{max} may approach the sublimation
temperature of the grain. Indeed as soon as the vapour pressure of the grain
materials becomes an appreciable fraction of the gas pressure, the grain will
be destroyed on a fairly rapid timescale.

12.4 Grain Charging

Advanced Topic
   First, let us consider charging through collisional processes only. The net
electric charge in the ISM is constrained to be zero. Additionally, at most
parts of the ISM we can consider that thermal equipartition is satisfied. This
implies that the electrons move faster than the ions in the velocity ratio
$v_e/v_i = \left(m_i/m_e\right)^{1/2}$, where m_i is the mass of the ions present. Thus, an
initially uncharged grain will collide with electrons at a greater rate than
with the ions, and will tend to pick up a net negative charge. The Coulomb
barrier imposed by this charge will then tend to decrease the collision rate
with electrons relative to the ions, until a current balance; a balance between
the rate of accumulation of negative and positive charge is achieved:

$$\int_{0}^{\infty} \sigma_e\left(v_e\right) n_e f\left(v_e\right) v_e dv_e = \int_{0}^{\infty} \sigma_i\left(v_i\right) n_i f\left(v_i\right) v_i dv_i \qquad (12.34)$$

where $f(v)$ is the Maxwellian distribution in velocity at the gas temperature T, and $\sigma_e(v_e)$, $\sigma_i(v_i)$ are the effecive grain cross - section to collisions with electrons and ions, respectively, taking into account the Coulomb interactions due to the grain charge $-Ze$:

$$\sigma_e(v_e) = \pi a^2 \left[1 - \frac{2Ze^2}{am_e v_e^2} \right]$$

$$\sigma_i(v_i) = \pi a^2 \left[1 + \frac{2Ze^2}{am_i v_i^2} \right] \tag{12.35}$$

These two equations have the solution:

$$1 - y = \left(\frac{m_i}{m_e} \right)^{1/2} \exp[y] \text{ with:} y = \frac{Ze^2}{akT} \tag{12.36}$$

In an ionised region, where the temperature is around 10000K and the ions are mostly protons, the collisional grain charge would be about $400e^-$ while in an H I region where temperature is only ~ 100K and the dominant ion is C II, the charge is only about $6e^-$.

When a sufficiently strong UV photon field is present, the photoelectric emission is the dominant grain charging process, and the grain charge is then positive. Photoelectrons produced by grains contribute to the photoelectric heating of ionised nebulae. Because at a given point in an H II or H I photodissociation region, the mean grain potential is often less than the mean ionisation potential of the ionic species present, this photoelectric heating effect can be quite important in determining the local electron temperature.

Considered purely as an absorber of EUV radiation, dust would act more as a cooling agent in ionised plasma, since the rather grey opacity of the grains at these wavelengths tends to produce a softening of the radiation field, and a lowering of the electron temperature (Petrosian *et al.*, 1972; Sarazin, 1977). However, when grain photoelectric heating is taken into account, dust acts as a powerful energy source in both H I regions (Draine 1978) and H II regions (Maciel & Pottasch 1982; Oliveira & Maciel, 1986).

As shown by Draine (1978), one of the important parameters governing the effective grain heating is the photoelectric yield as a function of the photon energy. This yield is characterised by two parameters, the limiting yield Y_∞ ($\sim 0.1 - 0.5$), and the threshold for photoelectron production, $E_{\min}$, typically 5 - 8 eV for most grain materials. As a function of photon energy $h\nu$:

$$Y(\nu) = Y_\infty(1 - E_{\min}/h\nu) \tag{12.37}$$

In laboratory experiments, a peaked energy distribution function of the emitted photoelectrons spectrum is obtained as a function of photon energy ($h\nu$). A reasonable approximation to this energy distribution function, $f(e, h\nu)$, is a triangular distribution:

$$e < (h\nu - E_{\min})/2 : f(e, h\nu) = \beta e/(h\nu - E_{\min})$$
$$e \geq (h\nu - E_{\min})/2 : f(e, h\nu) = \beta(1 - e/(h\nu - E_{\min})) \qquad (12.38)$$

where $\beta = 4/(h\nu - E_{\min})$.

In equilibrium, the collisional and photoelectric grain currents must balance (*see* Draine (1978). The photoelectric current, j_ν, due to the ambient UV radiation field, $U(\nu)$, is determined from the following integral:

$$j_\nu = \pi a^2 \int_{E_{\min}}^{\infty} \left[\int_{Max(E_g,0)}^{h\nu - E_{\min}} f(e, \nu) de \right] Q_{\text{abs.}}(\nu) Y(\nu) U(\nu) d\nu, \qquad (12.39)$$

where E_g is the grain potential, and $Q_{\text{abs.}}(\nu)$ is the grain absorption efficiency. This photoelectric current must be integrated numerically.

12.5 The Life Cycle of Grains

After being born as stardust, the interstellar grain population does not remain static with time. In dense interstellar clouds, dust grows volatile icy mantles consisting of water (H_2O), methanol (CH_3OH), carbon monoxide (CO), carbon dioxide (CO_2) or methane CH_4, many of which have been observed in the dense molecular gas towards the Galactic Centre. These ices may form more complex species through surface reactions, while other more refractory elements are captured and incorporated into the mantles. In the presence of a UV photon field, the volatile ices are photolysed into more stable organic grain materials (Greenberg, 1982), and in cool molecular regions, grains will coagulate together to form larger complexes of mixed composition.

Grain growth by atomic sticking has already been described in section (12.1). The characteristic timescale for significant loss of heavy atoms from the gas phase (the condensation timescale, τ_{cond}) is $\tau_{cond} \sim 10^{10} n^{-1}$ yr. Thus the heavy elements are depleted from the gas phase in a timescale of 10^6 or 10^7 yr in a typical molecular cloud. This process may be limited to some extent by desorption effects such as grain heating by cosmic rays and photo- and chemi- desorption.

In low density regions, grain destruction processes dominate the life cycle of grains. Supernova shocks, or shocks driven by the stellar winds of hot stars sweep through the low density gas, sputtering the grains if the velocity is high enough (Dwek *et al.* 1996), or shattering the grains at lower velocity (Jones *et al.* 1996). Close to hot stars, grain heating destroys mantles by sublimation, and the grains may even become charged to such a degree that the electrostatic repulsion exceeds the strength of the inter-atomic bonds and the grain is destroyed by cold field emission.

Very little of the current interstellar grain population can be the original stardust. For one thing, the sources of stardust are inadequate. In cool star winds, only the production of carbon is adequate to account for the current interstellar abundance. Other refractory elements such as Si, Mg and Fe must have been mostly made in supernova explosions. Although such explosions may produce dust in the fireball, this would have been almost completely destroyed during the passage through the reverse shock front. The mean timescale for destruction of dust by processing through shocks in the interstellar medium is only about 5×10^8 yr near the sun. The very high depletion factors which are observed for refractory elements (even along the lower-density lines of sight) require that the dust is re-formed on a timescale which is considerably shorter than this.

The evolution of the original stardust grain population and the new dust population formed in dense interstellar clouds is therefore a dynamic balance of all of these growth, destruction, structural and chemical modification processes, described in more detail in Section (14.3.5). Along any line of sight all these environmental factors determine the grain composition, the size distribution, the extinction law and the chemical depletion factors that will be observed.

12.5.1 Thermal Sputtering of Grains

Advanced Topic

Interstellar shocks from supernova remnants, or from fast winds of hot stars, are the most effective means of destroying interstellar grains. However, the dominant mode of destruction depends on the shock velocity. If the shock velocity is high enough ($\gtrsim 400$ kms^{-1}), then *sputtering* by fast grain-ion collisions will dominate, while at lower velocities grain-grain collisions will lead to grain *shattering* (see next section).

A grain will enter the shock moving at $3/4$ of the shock velocity with respect to the ions. Whether the grain survives the initial shock (assumed to be non-radiative because it is so fast) will depend on whether the grain can be slowed down by the drag forces of the surrounding gas before it is sublimated. In a medium of number density n, the drag force, F, acting on a grain of radius a and density ρ_g and moving with relative velocity v_g is;

$$F = \frac{4}{3}\pi a^3 \rho_g \frac{dv_g}{dt} = \pi a^2 \mu m_H n v_g^2 \tag{12.40}$$

where μ is the molecular weight and m_H is the mass of the hydrogen atom. Integrating this equation of motion from the point of entry of the grain into the shock until it slows to the characteristic thermal velocity of the surrounding medium gives the grain stopping timescale, τ_{stop} :

$$\tau_{stop} = \frac{4a\rho_g}{3\mu m_H n} \left(\frac{1}{v_{th}} - \frac{4}{3v_s} \right) \tag{12.41}$$

this is of order $10^6 a_{-5} v_{500}^{-1} n^{-1}$ years where a_{-5} is the grain radius in units of 10^{-5} cm and v_{500} is the shock velocity in units of 500 kms^{-1}. This stopping time will be modified somewhat by the effects of grain charging, and coupling of the grain to the magnetic field in the shock.

During the braking process the grain is heated by gas-grain collisions (*see* Dwek & Arendt, 1992). The smaller ($a \lesssim 0.05 \mu m$) grains will suffer subtantial temperature fluctuations, analogous to the quantum fluctuations discussed above (Draine & Anderson, 1985), and so the dust in the shock is an enhanced source of IR emission, particularly in the near IR.

The sputtering rate is a rather complex function of temperature, and threshold energy; given by Draine & Salpeter (1979). However, the results of Dwek *et al.* (1996) can be rather well fitted by an equation of the form:

$$\frac{da}{dt} = A n T_6^{-1/4} \exp \left[-B T_6^{-1/2} \right] \tag{12.42}$$

where T_6 is the gas temperature in units of 10^6K and A and B are constants. For graphite, $A = 6 \times 10^{-6} \mu m$ yr^{-1} and $B = 3.7$ while for silicate $A = 1.8 \times 10^{-5} \mu m$ yr^{-1} and $B = 4$. The sputtering lifetime for average size grains ($a \sim 0.1 \mu m$) is of order 10^5 years in fast shocks. This is comfortably shorter than the grain braking timescale, so that only the very largest grains can survive.

In fast shocks, the effect of the magnetic field can be important. Since the magnetic field is compressed in the cooling gas behind the shock front, charged grains are accelerated around the field lines in a "betatron acceleration" to produce gas-grain relative velocities comparable with the shock velocity (Spitzer, 1976). This is the main mechanism for grain destruction for J-shocks in the velocity range 50-200 km s^{-1} (Jones *et al.* 1994, and references therein). In C-shocks, it is the drift between the weakly-coupled ions and neutrals that provides large non-thermal streaming motions. Since the grains are charged, they tend to follow the motion of the ions, and the collisions between grains and neutrals produce non-thermal sputtering. This was computed by Pineau des Forêts & Flower (1997), who also calculate the sputtering yields.

12.5.2 Grain-Grain Collisions

Advanced Topic

Grain-grain collisions play a major role in determining the grain size distribution function. The effect of these collisions depends upon the impact velocity. At very high velocities, grain-grain collisions will drive a strong shock into each of the colliding particles, compressing them to high pressures and melting them. After the shock has passed through the grain and the pressure is relieved, the grains will vaporise.

At lower velocities when the velocity of impact is somewhat not too much greater than the typical sound speed in the grain (of order 5 km s^{-1}), the

shock wave transmitted into each grain has enough energy to shatter the rather brittle grains, producing many more smaller grains. At still lower velocities, below about 1 km s^{-1}, the grains will simply bounce off one another, and at the lowest energies, grain coagulation will occur.

The process of grain shattering in intermediate-velocity shocks has recently been investigated Jones *et al.* (1996). In essence, small fragments are formed as a result of cratering during the grain-grain impact. Matter ejected at different distances from the impact centre is ejected at different velocities, producing stresses in the ejecta which lead to fragmentation. The characteristic size of the ejected fragments is determined by matching these stresses to the shear strength of the grain material. The size distribution of the fragments is therefore determined by the shape of the cratering velocity field, and the ratio of the maximum to minimum fragment sizes is determined by the ratio of the pressure at impact compared to the pressure at which the flow ceases to be plastic. A somewhat different kind of fragmentation occurs when the shock remains strong in passage through the grain, as will tend to occur in the collision of two equal-sized grains and at higher velocities of impact. Here the compressive shock wave is reflected on the far side of the grain as a tensile wave, leading to failure under tensile stresses (*spallation*). This produces a shower of small fragments, and may result in the complete break-up of the grain. For typical grain materials, the onset of fragmentation by cratering is of order 2 km s^{-1} and catastrophic spallation occurs at 75 km s^{-1} for amorphous carbon, and at 175 km s^{-1} for silicates.

This shattering process leads naturally to a power-law size distribution of grains; $dn/da = ca^{-\alpha}$ with $\alpha \sim 3.3$, observationally indistinguishable from that adopted by Mathis, Rumpl, & Nordsieck (1977); $\alpha \sim 3.5$. Essentially, any shattering model in which the size of the fragments is related to the pressure experienced in the collision will lead to a power-law distribution with a slope somewhat greater than 3.

Thus, lower-velocity shocks are a copious source of small grains, and may even provide the main source of the small aromatic grains and molecules need to explain the PAH features seen in the infra-red between 6 and 12μm.

During a collision, the surfaces of the grains elastically deform and store the energy needed to cause the particles to bounce off one another. However, during a low energy collision, some of this energy is lost by the excitation of lattice phonons, and the collision is not perfectly elastic. When the internal energy stored as phonons is comparable with the grain-grain collisional energy, the particles will stick together. In a sense, therefore, this process of *grain coagulation* is similar to the process of atomic sticking, discussed in section (12.1). The maximum velocity for coagulation, v_{coag}, (typically in the range 10^2–10^4 cm s^{-1}) is given by (Tielens, 1989):

$$v_{coag} = \left(\frac{E_{int}}{Y a_{red}} \right)^{5/6} c_g$$

where E_{int} is the interface energy, Y the Young's Modulus of the grain material with sound speed c_g, and a_{red} is the reduced radius of the colliding grains; $a_{red} = a_1 a_2 / (a_1 + a_2)$. Grain coagulation is therefore assisted in small grains, and in grain materials with low Y which are easily deformable, such as ices. In dense molecular clouds, the low temperatures and the formation of icy mantles will assist coagulation. Since the grain-grain collision timescale is also considerably shorter than the cloud lifetime, large complexes of coagulated grains may be built up.

12.6 Spectral Signatures of Grain Materials

Molecular bonds have characteristic frequencies for stretching, bending, or scissoring modes. When incorporated into grain materials, these characteristic frequencies are shifted somewhat and broadened by the effects of the lattice. Nonetheless, there are many prominent absorption or emission features, particularly in the far-IR which have enabled us to obtain unamiguous identifications of a variety of interstellar grain types.

Combining data from satellites such as IRAS (Infra-Red Astronomical Satellite), COBE (Cosmic Background Explorer) and ISO (Infrared Space Observatory) with ground-based data and data obtained in the UV from IUE (International Ultraviolet Explorer) and the Hubble Space Telecope it has been possible to cover the whole dust spectrum. This provides new powerful observational constraints on the nature, composition and size distribution of the various dust grain species.

Research in this field is also assisted by the study of *interplanetary dust particles.* These can be collected from the earth's stratosphere using high-flying aircraft. Provided that the particle is small enough, it can be slowed by dynamical drag in the outer reaches of the earth's atmosphere without being melted, and subsequently gently drifts down and settles out. Such particles provide a useful source of dust materials which may be truly representative of interstellar dust types.

In the following subsections, we briefly review the key identifications which have been made in this field.

12.6.1 Constraints from the Spectral Energy Distribution

A successful grain model should be able to describe both the extinction and emission properties of the dust from the Lyman limit at 0.0912μmall the way up to 1000μm. The visible and ultraviolet absorption properties were discussed above. These constrain the size distribution of the grains, which is usually described by a power-law distribution $dn/da = ca^{-\alpha}$. The power-law index adopted is usually that of Mathis, Rumpl, & Nordsieck; $\alpha \sim 3.5$. The 2175Å bump demands the presence of small carbon grains in the form of

graphite with a size of about 0.02μm. However, as we have seen, this bump is variable in width, from about 350 to 600Å, and is correlated with the rise in the far-UV extinction and the value of R, showing that in dense clouds grain coagulation and mantle growth combine to change the size distribution, particularly at the small particle end.

A number of recent attempts have been made to model both the extinction and the continuum *spectral energy distribution* (SED) of dust emission in the IR (*see* Dwek *et al.* 1997, and references therein). The SED has a broad maximum at around 100-200μm, which can be ascribed to the larger silicate grains which are responsible for much of the optical extinction. The SED falls much less steeply than a black-body distribution towards shorter wavelengths. This requires a population of small grains, either of *small organic grains* (SOGs) or of small silicate grains. In addition, the 3 - 12μmregion is dominated by both line and continuum emission from the ubiquitous population of the polycyclic aromatic hydrocarbon (PAH) grain materials.

12.6.2 Silicates

Advanced Topic

The characteristic interstellar absorption features due to silicates are found at 9.7μmand at 18μm. These features are rather broad, the 9.7μmfeature ranging from a wavelengthof roughly 8μmup to 12.5μm. The 18μmfeature is shallower, and ranges from 15–22μm, approximately. These features have been seen in emission in the spectra of comets, such as Halley and Kohoutek or in the circumstellar material around the star β Pictoris.

The ubiquitous 10μmfeature arises from the Si–O stretching mode, and the 18μmfrom the Si–O–Si bending mode. These are prominent in olivine, but have a more complex structure in pyroxenes, because of the greater compositional complexity that is possible in these. Crystalline lattice-layer silicates additionally show bands at 3.1μmand 6μm. These are due to the O–H stretching and H–O–H bending modes caused by adsorbed and absorbed water. In addition, features at 6.9μmand 11.4μm are due respectively to the stretching and scissoring modes of carbonates (Sandford, 1989).

The emission spectrum of comets can be fit by a combination of these types; about 55% olivine, 35% pyroxene and 10% of lattice-layer silicates (Bregman *et al.* 1987). The best fit to the observed spectra is obtained if the grains are assumed to be coated in a carbon-based material.

12.6.3 Icy Grain Mantles

Advanced Topic

The 2-20μmspectra of objects lying behind dense molecular clouds frequently show deep absorption features at 3.08μm4.67μm, 6.0μm and 6.85μmalong with the well-known 10μmfeature due to silicates. In addition, high signal to

noise spectra taken with ISO reveal a number of weaker features in this wavelength region. Many of these features are due to interstellar ice mantles on grain surfaces, and their identifications have recently been reviewed by Tielens & Whittet (1997). In particular, ices due to CO_2 H_2O CH_4 and CH_3OH have been definitely identified, as well as an unidentified nitrile or isonitrile compound dubbed XCN.

The 3.08μmis caused by the O–H stretching mode in water ices. The broad 6.0μmfeature is due to water ices, a result of the O–H bending mode (Tielens *et al.* 1984). This is confirmed by the existence in ISO spectra of a weak H_2O absorption on the long wavelength side of the strong 10μmsilicate band.

The strong and narrow 4.67μmfeature is the signature of solid CO (Lacy *et al.* 1984). The central wavelength and shape of this feature depend on the nature and composition of the surrounding matrix. In the source W33, a strong satellite band is seen at 4.62μm, which may be the result of the C≡N stretching mode in nitriles or isonitriles produced by photolysis. An alternative explanation, that it is the Si–H stretching vibration in organic-silicon compounds (Nuth & Moore, 1988). If so, these too would have to be the result of UV photolysis in icy mantles containing silicon compounds such as SiH_4.

The identification of the 6.85μmfeature is plausibly the result of the C–H and O–H deformation mode in CH_3OH. However, methanol ices are certainly detected through their C–H stretching mode at 3.54μmand an overtone band at 2.4μm.

12.6.4 Polycyclic Aromatic Hydrocarbons (PAHs)

Advanced Topic

The recently obtained ISO spectra show that the emission features attributed to PAHs are ubiquitous in the interstellar medium, provided the line of sight probed is not too dense, or too highly excited. PAHs or other organic compounds containing aromatic ring molecules produce characteristic emission features at 3.28μm, 6.2μm, 7.6-8.0μm, 8.6μmand 11.3μm. Each of these features has been identified with a specific excitation mode of aromatic compounds. The 3.28μmfeature is the fundamental aromatic C–H stretching mode, the 6.2μmband is the aromatic C–C stretching mode, while the 7.6-8.0μmband is a complex blend of several excited aromatic C–C stretching bands. The longer wavelength features are due to lower-energy bending modes; the 8.6μmband results from aromatic C–H in-plane bending mode, and the 11.3μmband is caused by the aromatic C–H out–of-plane bending mode for non-adjacent peripheral H atoms (Allamandola *et al.* 1989).

Along different sight lines the carbonaceous grain types may be either predominantly aliphatic, with high concentrations of O and H or aromatic with low H and O concentrations, as evidenced by the detailed structure of the spectrum in in the 3.2 - 3.6μm region (Sauvage, 1998). The Galactic Centre source, Sag A, shows predominantly aliphatic carbon types, with a

primary carbon absorption peak at 3.4μm, the emission spectra of post-AGB stars show a mixture of types, and the reflection nebula NGC2023 shows mainly the 3.28μm PAH feature.

It is likely that UV photolysis will cause an evolution from PAH-like grains to graphitic types, as small PAH molecules are likely to lose their peripheral hydrogen atoms if sufficiently excited, leaving their carbon skeletons more or less intact. UV excitation of small PAHs (~ 60 atoms) is also though to produce the red and near IR fluorescent features seen in reflection nebulae (Sellgren *et al.* 1985), and a dissociation continuum (d'Hendecourt *et al.* 1989). In addition, IR emission in the $> 20\mu$m region can be induced by the excitation of C–C –C out-of-plane bending modes in the aromatic rings. The intensity of the IR emission is greatly assisted by the quantum heating effects described in section (12.3.1), which heats these molecules to temperatures of order 600K.

Additional constraints on the carbon grain types can be obtained by considering the total carbon abundance budget (Sauvage, 1998). In the ISM, there is about 200 parts per million (ppM) of carbon which is available to form dust. Of this, at least 80 ppM is required to fit the 2175Å bump. Most models cannot produce enough absorption and emission with what is left, as they require anywhere between 240 and 500 ppM of carbon, depending on the model. Although "fluffy" grain models can work within the carbon budget, these have problems fitting the absorption and IR emission constraints. This carbon problem is not yet fully resolved.

12.6.5 The Diffuse Interstellar Bands (DIBs).

Advanced Topic

We have known about the problem of diffuse band absorption in the spectrum of stars for over 50 years. Although we are confident that they have an interstellar origin (since they all correlate to some extent with the line of sight reddening), we still have only vague ideas about their origin.

The first band to be discovered was the band at 4430Å, but now nearly 200 bands are known between 4200Å and 8700Å (Jenniskens & Désert, 1995). The equivalent widths of the detected features (measured for $E(B - V) = 1$) range from a few Å down to about 0.01Å. In general, the broader the feature, the greater its equivalent width, although many more narrow bands are currently known because these show better contrast against the backround continuum source. Generally, the bands seem to fit into three loose "families" with representative members being the 5797Å, 5870Å and 4430Å bands (*i.e.* Krełowski & Walker 1987). Within these families, the lines show better correlations with one another. However, only the 5780Å and the 6284Å features show such a close correlation that we can be confident that they arise in the same carrier. Many of the features appear to be associated with the diffuse interstellar gas, rather than with dense clouds, since their correlation with the line of sight $E(B - V)$ is better than their correlation with molecular

hydrogen column density. Furthermore, except for the 4430Å feature, they become relatively weaker as the cloud density rises.

It now seems much more likely that they arise from a complex molecular species, rather than from the solid phase. This idea gathers support from the observations of the "red rectangle" by Miles *et al.* (1995). They found a number of bands in emission, including the 5797Å feature. The regular interval between the features, the blue asymmetry and shift of the peak wavelength as a function of emission strength all suggest that a series of vibrational states of a complex molecule is being observed. Additionally, very high resolution observations of the 6614Å feature by Hibbins *et al.* (1995) shows some evidence for a Q, P and returning R branch of an unknown molecule.

Two intriguing ideas for the origin of these features have been proposed. One suggestion is that they arise in PAH^+ ions which originate from grain fragmentation in shocks. Another proposal holds that they come from fullerene "bucky-balls" of carbon (C_{50}, C_{60}, C_{70} *etc.*) with hydrogen atoms bonded to their outer surfaces, thus converting them to fulleranes. A complete solution to the problem of the diffuse interstellar bands still seems to be a long way off.

Notes on Chapter 11

- A fuller account of a number of the topics covered in this chapter is given in Spitzer, L., Jr., 1977 *Physical Processes in the Interstellar Medium* (Wiley:New York), ISBN 0-471-02232-2, now republished in the Wiley Classics Library Series, Wiley:NY.
- An excellent description of the formation of dust, and dust chemistry can be found in *Dust and Chemistry in Astronomy*, eds T.J. Millar and D.A. Williams, in The Graduate Series in Astronomy, Series Editors R.J. Tayler and R.E. White, Institute of Physics Publishing:Bristol and Philadelphia, ISBN 0 7503 0271 2.
- For an up to date account of the diffuse interstellar bands, see *The Diffuse Interstellar Bands*, eds. A.G.G.M. Tielens & T.P. Snow Klwer: Dordrecht, 1995. ISBN 0-7923-3629-1.
- A fine general review of circumstellar chemistry is by A. Omont, 1991, in *Chemistry in Space*, eds. J.M. Greenberg & V. Pirronello, Kluwer: Dordrecht, p171.

Exercises

Exercise 12.6.1. The following table gives the (absolute) relative abundances of the elements in the sun ($\log_{10}$ by number of atoms with respect to hydrogen). Assuming that the end-products of silicate condensation are, in order of condensation, melilite, $Ca_2Al_2SiO_7$ and diopside: $CaMgSi_2O_6$ and the alkali alumino-silicates: $NaAlSi_3O_8$ and, for the excess Mg and Fe, enstatite $MgSiO_3$ and olivine $(Mg,Fe)_2SiO_4$.

(a) Assuming that helium has an abundance 1/10th that of hydrogen, calculate the maximum mass fraction of silicate grains that can be formed, and the gas-phase depletion of oxygen expected when these grains have condensed out

(b) In order to explain the optical extinction curve, we would require a silicate mass fraction of 9×10^{-3}. Is the value you calculated in (a) sufficient, and if not, what other grain types can you suggest may be responsible, and why?

C	-3.44	Ne	-3.91	Al	-5.53	Ca	-5.64
N	-3.95	Mg	-4.42	Si	-4.45	Fe	-4.33
O	-3.07	Na		S	-4.79	Ni	-5.75

Exercise 12.6.2. Prove equation (12.36).

13. Introduction to Astrochemistry

'If you open that Pandora's Box, you never know
what Trojan horses will jump out!"
— Ernest Bevin

At first sight, the interstellar medium presents a very hostile environment to the formation and survival of complex molecules. As we have seen, it is criss-crossed by violent shocks and pervaded by intense UV and X-ray radiation fields, which can easily tear apart the delicate molecular bonds. Even the composition of the interstellar medium presents a problem to the formation of complex molecules, since the gas is mostly hydrogen and helium. Reactive species are simply trace elements, and direct collisions between them rare. Nonetheless, astrochemistry turns out to be a rich and complex field. In cool regions in the tails of shocks gas phase reactions can occur on interestingly short timescales. Within dense clouds, UV photons and cosmic rays help moderate a whole series of chemical reactions. On the surfaces of grains, migrating molecules can meet and react together, and very complex molecular species can be built up. In this Chapter, we will simply get a taste for these phenomena. To find out more the recent book by Emma Bakes (1997), and the reviews by Dalgarno (1987) and Genzel (1992) are strongly recommended (*see* notes on this chapter).

13.1 Molecular Formation and Destruction

13.1.1 Neutral Gas-Phase Reactions

The simplest type of reaction that we can think of is simply a "sticking" collision between a pair of neutral species. This is called the two-body *radiative association*:

$$A + B \rightarrow AB + h\nu \qquad (13.1)$$

For such a reaction to proceed, the radiative processes have to be very fast, as the interaction lasts only for a collisional timescale $\tau_{coll} \sim 10^{-8}/\langle v_{therm} \rangle$ s, where $\langle v_{therm} \rangle$ is the mean thermal velocity. For favourable conditions (in a dense cloud with temperature ~ 10 K), this is less than 10^{-12} s. However, as we have seen, an allowed dipole transition may have a transition probability

of order 10^8 s^{-1}. Thus, only about one collision in 10^4 to 10^5 will induce a reaction, and radiative association is therefore very rare.

Chemical reactions between neutral particles may also occur. These usually have an activation energy barrier to overcome, so that the association reaction is endothermic. Such neutral-neutral reactions are therefore very unlikely at the low temperatures of molecular clouds. However, there are some exceptions, such as the reaction: $O + OH \rightarrow O_2 + H$, in which the approach can take place along a potential energy surface with a deep potential well.

If the gas is warmed by an external means, neutral-neutral reactions may overcome the energy barrier, and these can then become important coolants for the gas. Examples of such warm environments include protostellar or evolved star winds, the warm regions of a C-shock, or the zone following the recombination region in a faster J-type shock. The types of neutral reactions that are most important are:

$$A + H_2 \rightleftharpoons AH + H \qquad (13.2)$$

and:

$$AH + B \rightarrow AB + H \qquad (13.3)$$

where A, B are O, C. N, S or Si. The rate coefficients of such reactions can be fitted as a function of temperature by an expression of the form:

$$R = \alpha T_{100}^{\beta} \exp\left[-\gamma/T\right] \text{cm}^3 \text{ s}^{-1} \qquad (13.4)$$

where α is a constant, T_{100} is the temperature in units of 100K, and γ is the activation temperature of the reaction (K). Rate coefficients of these and several thousand other reactions are maintained in the UMIST database for astrochemistry (Millar *et al.* 1997) which represents the major source of such data (see notes).

The relatively small energy barrier (3940K) for the $O + H_2 \rightarrow OH + H$ reaction coupled with the large oxygen abundance ensures that OH is by far the most abundant hydride produced by reactions of the type (13.2). The corresponding reaction with C has a much higher energy barrier (14100K), so is less likely to occur. The OH and CH which are formed by these endothermic reactions can be transformed to O_2, CO and C_2 by neutral reactions with O or C atoms.

13.1.2 Ion-Molecule Chemistry

When *molecular ions* can be formed by cosmic ray ionisation or by photoionisation in an interstellar cloud, a large number of ion-molecule reactions can be activated. Unlike the neutral-neutral reactions these remain fast at low temperatures, and so come to dominate the chemistry at low temperatures. The ion in its approach to the neutral polarises the charge cloud around the neutral to produce an attractive potential. Because the interaction timescale

is long at the low temperatures appropriate to molecular clouds, the reaction rates are then limited by the *Langevin* rate coefficient appropriate to adiabatic collisions (*c.f.* Section (5.4)):

$$R = 2\pi e \left(\frac{\alpha}{\mu}\right)^{1/2} \sim 10^{-9} \text{cm}^3 \text{ s}^{-1} \tag{13.5}$$

where α is the polarisability and μ the reduced mass of the reactants. For typical molecular clouds, the reaction timescales will be typically a few thousand to a few million years, depending on the ionic concentration.

For longer-range collisions involving heteronuclear molecules, in addition to the polarisation attraction, there is also an attractive force caused by the permanent dipole of the heteronuclear molecule. Depending on orientation of the collision, the reaction probability may be increased. This can lead to a total reaction rate and collision frequency which actually increase with decreasing temperature.

At very low temperatures, the relative populations in the fine structure levels change. For example in reactions involving atomic oxygen, the atoms are restricted to the $^3P_2$ level at sufficiently low temperatures, resulting in quite large changes in reaction rates.

While *radiative association* reactions are usually slow for neutrals, they are important in ion-molecule reactions. For example, in photodissociation and cosmic-ray ionised regions, the carbon chemistry is initiated by the radiative association reaction:

$$C^+ + H_2 \rightarrow CH_2^+ + h\nu \tag{13.6}$$

and quite large molecules can be formed in this way, for example;

$$CH_3^+ + H_2O \rightarrow CH_3OHH^+ + h\nu \tag{13.7}$$
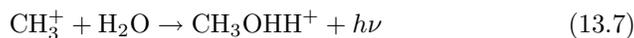
Complex neutral species can be built up by a series of *hydrogen abstraction* reactions of the type:

$$AH_{n-1}^+ + H_2 \rightarrow AH_n^+ + H \tag{13.8}$$
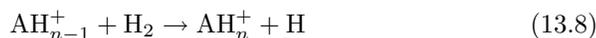
followed by a *dissociative recombination*:

$$AH_n^+ + e \rightarrow AH_{n-1} + H \tag{13.9}$$

13.1.3 Dust-Grain Moderated Chemistry

In cool clouds the dust grains form excellent sites for chemical reactions because reactive molecular or atomic species may be first adsorbed onto the grain surface and then migrate around the surface (*see* Section (12.1)), until they meet other species with which they can combine in order to form a

new molecule. Depending on how exothermic the chemical reaction is, the species that is produced in this way may either be desorbed from the surface, or remain to be chemically incorporated into the icy mantle of the grain. Because the timescale of the encounter on the dust grain surface is very long, even reactions which are most unlikely to occur in the gas phase are enabled. In this sense, grain surfaces act as a *catalyst* for such reactions.

The reactivity of carbon surfaces is enhanced by their microcrystallinity. This produces chemically active sites due to unoccupied states in the outermost atoms which are available to bond with other atoms or molecules. Such sites are called *dangling bonds* or *unsaturated valencies.*

If the probability (or yield) Y, of species A being adsorbed following a collision with a grain, migrating and reacting with species B to form the molecule AB, then the rate of this molecular formation reaction would be:

$$R_{AB} = Y \sigma_A v_A n_A n_{gr} \mathrm{cm}^{-3} \ \mathrm{s}^{-1} \tag{13.10}$$

here, n_A and n_{gr} are the number densities of the species A and the grains, respectively. The thermal speed of the species A is $v_A = (8kT/\pi m_A)^{1/2}$ and σ_A is the effective cross - section of the grain, taking into account the effect of the charge on the species A (if any) and the grain charge, according to equation (12.35). Since the grains tend to have a net negative charge in dense clouds due to collisional charging, and the ions have positive charge, the collision rate will tend to be enhanced over that estimated by taking only the geometrical considerations into account. The effective yield can be quite high, $0.4 < Y < 0.8$; *see* Section (12.1). The reaction timescale is therefore:

$$\tau \sim 2.6 \mathrm{x} 10^7 \frac{\mu_A^{1/2}}{Z(A) Y T_{10}^{1/2} a_{-6}} n_H^{-1} \mathrm{yr} \tag{13.11}$$

where n_H is the hydrogen number density of the cloud, μ_A is the molecular or atomic weight of species A, $Z(A)$ is its abundance by number relative to molecular hydrogen, a_{-6} is the mean effective grain radius, normalised to 10^{-6} cm, and T_{10} is the cloud temperature in units of 10K.

One of the major species that must be formed on the surfaces of grains is molecular hydrogen, since possible gas-phase reactions proceed far too slowly to explain its abundance. For this species, Hollenbach & Salpeter (1971) estimate that the yield is large, and if this is the case then dust grain surface reactions can form this species within a small fraction of the lifetime of a typical molecular cloud ($\tau_{cloud} \lesssim 10^7$ yr). The exact mechanism of formation remains uncertain, however. These processes are reviewed by Williams (1987).

Other molecular species are formed on longer timescales, and heavy atoms can be depleted down to a level of about $Z(A) \sim 5 \mathrm{x} 10^7 \tau_{cloud}^{-1} n_H^{-1}$ in the cloud lifetime, unless molecules are returned to the gas phase by other processes such as *chemical desorption*, through internal molecular cloud shocks (which induce *thermal desorption* and/or *sputtering*), or by desorption induced by photons (Boland & de Jong, 1982) or by cosmic rays (Lëger *et al.* 1985).

Photons may also assist the formation of molecules on grain surfaces. For example H_2CO and CO_2 can both be formed by photon-moderated grain surface reactions at a rate coefficient of about 5×10^{-12} cm^3 s^{-1};

$$CO + H_2O + h\nu \rightarrow CO_2 + 2H$$
$$\rightarrow H_2CO + O \qquad (13.12)$$
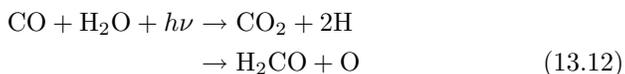
In the reaches of interstellar space, the icy mantles of grains remain cool, and the bombardment of far-UV photons creates radicals which are stable at such low temperatures. When the grain enters a hotter environment, these radicals react to create complex organic molecules such as HNC_3 or C_9N. Even more complex non-volatile materials can be created by photolytic reactions, including such species as complex polymers, glycine or amino acids - the very stuff of life!

The photolytic reactions induced by UV photons store energy in the mantles of grains in the form of chemical bonds. In shocks, chemically-induced grain mantle explosions can occur when a grain-grain collision heats the mantle to a sufficient temperature to suddenly release all this stored energy (d'Hendecourt *et al.* , 1982).

13.1.4 Molecular Destruction Processes

It may seem obvious, but is worth stating explicitly, that molecules can be destroyed wherever there exists a process to input more energy than the chemical energy associated with the molecular bond. This requires either heat input (through energetic thermal particle impact - usually electrons or through cosmic ray impact) or energy input by light in the form of UV photons. Let us consider photon-induced reactions. A typical *photodissociating* reaction is:

$$AB + h\nu \rightarrow A + B \qquad (13.13)$$

Because the scattering function of grains is strongly peaked in the forward direction, this field may be considerably stronger deep within the cloud than would be naively expected on the basis of extinction optical depth estimates.

A very important photodissociation reaction applies to CO. In this particular case the photon first excites discrete predissociating bound states:

$$CO + h\nu \rightarrow CO^* \rightarrow C + O \qquad (13.14)$$

There are about 44 bands due to pre-dissociated states in the 912-1100Å range (Viala *et al.* 1988; van Dishoeck & Black 1988). Since UV photons in these bands are used up by each excitation into a pre-dissociating state, this implies that CO can be self-shielding (Lee *et al.* 1996 and references therein). That is to say, if there is enough of a CO column density between the source of UV photons and the CO molecule considered, the absorption of the UV field by photoexcitation into these states will decrease the flux at these frequencies

sufficient to protect the CO molecule from photodissociation. The decrease in UV photons through self-shielding is much more important than the dust extinction in determining the photodissociation rates.

For radio observations of CO clouds, when both the ^{12}CO and ^{13}CO lines are optically-thick, the emission line flux ratio from an external galaxy simply depends on the ratio of the areas of optically-thick emission presented to the observer, since the total flux is the area times the linewidth times the Black-body function at the frequency concerned, see section (4.1.3). Where there is a strong UV-field, the ^{13}CO will be photodissociated to greater depth in the molecular cloud than the ^{12}CO as a result of this self-shielding. This probably explains why the ^{13}CO/^{12}CO emission line ratio is observed to be systematically lower in the most luminous starburst galaxies (Taniguchi & Ohyama, 1998).

The photodissociation of molecular hydrogen is also controlled by self-shielding. Far-UV photons are absorbed to excite the Lyman and Werner transitions in the 912-1100Å range. Photodissociation results from fluorescence to the vibrational continuum of the ground electronic state with a probability of about 0.1-0.15 per transition. Self-shielding becomes important once the H$_2$ column exceeds about 10^{14} cm^{-2}. Since self-shielding is much more effective in this abundant species, H$_2$ is much less likely than CO to be photodissociated by UV photons below the Lyman limit, so an absence of CO within a molecular cloud should not be construed as implying an absence of H$_2$, as was clearly demonstrated by Maloney & Black (1988). The importance of self-shielding in controlling the CO abundance at low C and O abundance is graphically illustrated in the case of the Large Magellanic Cloud (LMC). Although the oxygen abundance is only a factor of two lower than in the interstellar medium in the solar vicinity, the observed intensity ratio $I_{\mathrm{CO}(1-0)}/I_{[\mathrm{C\,ii}]}$ is a factor of 20 times weaker in the LMC than in the solar vicinity (Mochizuki *et al.* 1994). Since (*see* Section (4.1.3)) for the optically-thick transition such as CO (1-0), the line flux is proportional to product of the black-body function at this wavelength, the line width and the surface area of the clouds containing CO, this result shows that a much greater proportion of the molecular clouds are occupied by photodissociation regions in the LMC.

Photons may also induce *dissociative photoionisation* reactions of the form:

$$\mathrm{AB} + h\nu \rightarrow \mathrm{A} + \mathrm{B}^+ + e^- \tag{13.15}$$
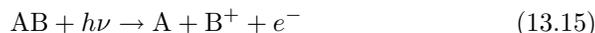
or else simple *photoionisation* reactions:

$$\mathrm{AB} + h\nu \rightarrow \mathrm{AB}^+ + e^- \tag{13.16}$$

Both of these input free electrons into the plasma. Electrons may also be generated in predominantly neutral gas through comic-ray induced ionisations:

$$\mathrm{AB} \overset{CR}{\rightarrow} \mathrm{AB}^+ + e^- \tag{13.17}$$

such reactions mostly ionise H_2 to H_2^+ in molecular clouds, since this is by far the most abundant molecule. The H_2^+ ion then reacts with a second H_2 molecule to form the H_3^+ ion, which is the key ion moderating dense cloud chemistry.

Free electrons provide cooling to molecular regions by exciting the electronic, rotational and vibrational states of molecules. However, they can also provide an important vehicle for destroying molecules. The simplest such reaction is *collisional dissociation*:

$$AB + e^- \rightarrow A + B^* + e^- \tag{13.18}$$
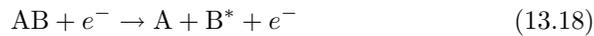
Electrons may also produce *dissociative ionisation*:

$$AB + e^- \rightarrow A + B^- + 2e^- \tag{13.19}$$
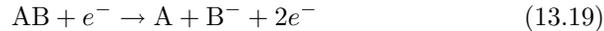
or (less likely) *dissociative attachment* reactions:

$$AB + e^- \rightarrow A + B^- \tag{13.20}$$

However, they are more likely to destroy positive ions produced by photoionisation or cosmic ray ionisation reactions and the subsequent ion-molecule chemistry through *dissociative recombination* reactions of the form:

$$AB^+ + e^- \rightarrow A + B \tag{13.21}$$

13.2 Chemistry of Particular Regions

13.2.1 Cold Molecular Clouds

Cold clouds such as the Taurus Molecular Cloud 1 (TMC-1) are very rich in complex linear unsaturated molecules. They have temperatures of order 10K, are very dense, $n\,(H_2) \sim 10^4$ cm^{-3}, and have very large dust optical depths; $A_V \sim 10$ mag. In such dense clouds the effects of photon-induced chemical reactions can be neglected. The primary source of ionisation is through cosmic rays; *see* (13.17) which provide $\sim 10^{-17}$ ionisations per H atom per second, sufficient to maintain a fractional ionisation of about $\sim 10^{-8}$. Cloud heating is from both cosmic rays and from turbulence, often associated with the formation of low-mass stars. Modern quantitative models for interstellar chemistry involve computation of reaction networks containing of order one thousand coupled, time-dependent, non-linear kinetic equations. Clearly, this is far beyond the scope of this book, so here we will consider only a few of the most important reactions that are involved.

The key molecular ion for interstellar chemistry is the production of the H_3^+ ion by *cosmic ray induced ionisation* of H_2, as was first shown by Herbst & Klemperer (1973). This is important right down to the hearts of even

the densest molecular clouds, since the stopping column for energetic cosmic rays is much greater than the cloud column density and the main energy loss mechanism is through knock-on ionisations along the cosmic ray track. The molecular ion H_2^+ is produced by the initial cosmic ray ionisation, but this quickly reacts through the ion-molecule process:

$$H_2^+ + H_2 \rightarrow H_3^+ + H \tag{13.22}$$

which opens the way for a rich chemistry involving proton transfer reactions of the type:

$$H_3^+ + A \rightarrow AH^+ + H_2 \tag{13.23}$$

Although the H_3^+ ion is of pivotal importance in molecular cloud chemistry, it sucessfully eluded detection for many years. Howerver, recently it has been finally identified by Geballe & Oka (1996) through a transition at $3.668\mu m$at a telative abundance $H_3^+/H_2 \sim 2 \times 10^{-9}$, which is consistent with the hypothesis that it is produced by cosmic ray ionisation.

An illustrative example of the way in which the H_3^+ ion helps form the very abundant molecules in dense molecular clouds is provided by the OH_n chain reactions, which dominate at cloud densities less than about $100 cm^{-3}$:

$$
\begin{aligned}
H_3^+ + O \quad &\rightarrow OH^+ + H_2 \\
OH^+ + H_2 \quad &\rightarrow OH_2^+ + H \\
OH_2^+ + H_2 \quad &\rightarrow OH_3^+ + H \\
OH_3^+ + C \quad &\rightarrow H_2CO^+ + H \\
H_2CO^+ + e \quad &\rightarrow CO + 2H
\end{aligned}
\tag{13.24}
$$
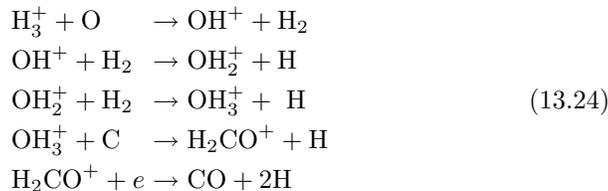
Here the final product is CO. Alternatively, the OH_3^+ ion may make a dissociative recombination to form OH:

$$OH_3^+ + e \rightarrow OH + 2H \tag{13.25}$$

Carbon monoxide can also be made the corresponding CH_n chain reactions, which dominate at cloud densities above a few hundred cm^{-3}:

$$
\begin{aligned}
H_3^+ + C \quad &\rightarrow CH^+ + H_2 \\
CH^+ + H_2 \quad &\rightarrow CH_2^+ + H \\
CH_2^+ + H_2 \quad &\rightarrow CH_3^+ + H \\
CH_3^+ + O \quad &\rightarrow H_2CO^+ + H \\
H_2CO^+ + e \quad &\rightarrow CO + 2H
\end{aligned}
\tag{13.26}
$$
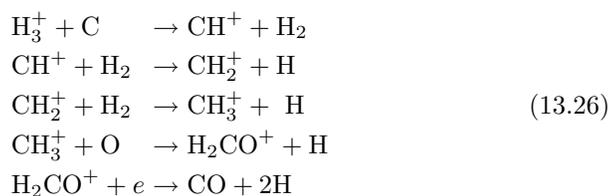
Here, the CH_3^+ ion can also undergo dissociative recombination to produce CH:

$$CH_3^+ + e \rightarrow CH + 2H \tag{13.27}$$

Depending on the relative abundances of the corresponding atomic species, the CH_3^+ ion that is formed in the CH_n chain may react instead with C to form C_2:

$$CH_3^+ + C \rightarrow C_2H_2^+ + H$$
$$C_2H_2^+ + e \rightarrow C_2 + 2H \tag{13.28}$$

or with N to form CN:

$$CH_3^+ + N \rightarrow HCN^+ + 2H$$
$$HCN^+ + e \rightarrow CN + H \tag{13.29}$$
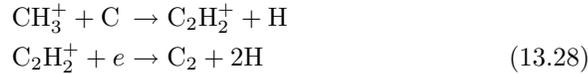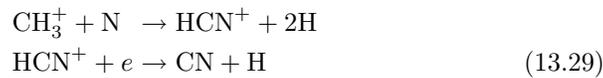
A more direct way of making C_2 is through the ion molecule reaction:

$$CH + C^+ \rightarrow C_2^+ + H \tag{13.30}$$

followed by the charge-exchange reaction:

$$C_2^+ + C \rightarrow C_2 + C^+ \tag{13.31}$$
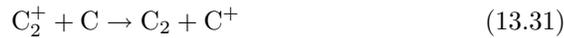
Alternatively the C_2^+ ion passes through the ion-molecule reaction:

$$C_2^+ + H_2 \rightarrow C_2H^+ + H \tag{13.32}$$

followed by dissociative recombination:

$$C_2H^+ + e \rightarrow C_2 + H \tag{13.33}$$

13.2.2 Photodissociation Regions

Beyond the ionisation fronts of H II regions in the region where all the photons with energies above the Lyman limit have been used up in photoionisation, there is still a very strong UV radiation field below the Lyman limit which provides the dominant heating source for the gas. Initially, the photon field is absorbed by the atomic species with ionisation potentials below 13.6 eV; C^+, S^+, Si^+, Mg^+ and Fe^+ are the dominant ionic species here, and the main atomic cooling process occurs through the O I fine structure line at $63\mu m$ and the C II fine structure line at $158\mu m$. However, dust is an important source of opacity for these UV photons, and photoelectric heating of the gas may be the dominant heating term, depending on the nature and size distribution of the dust grains; *see* Section (12.4). This heating is dominated by the smallest grains present.

Deeper within the cloud, molecules can be found. Since the ionisation front is evolving towards this molecular gas and the far-UV radiation field is therefore increasing with time, the molecules are being destroyed as they move through this *photodissociation region*. The structure of such regions was worked out in detail by Tielens & Hollenbach (1985), and the field was
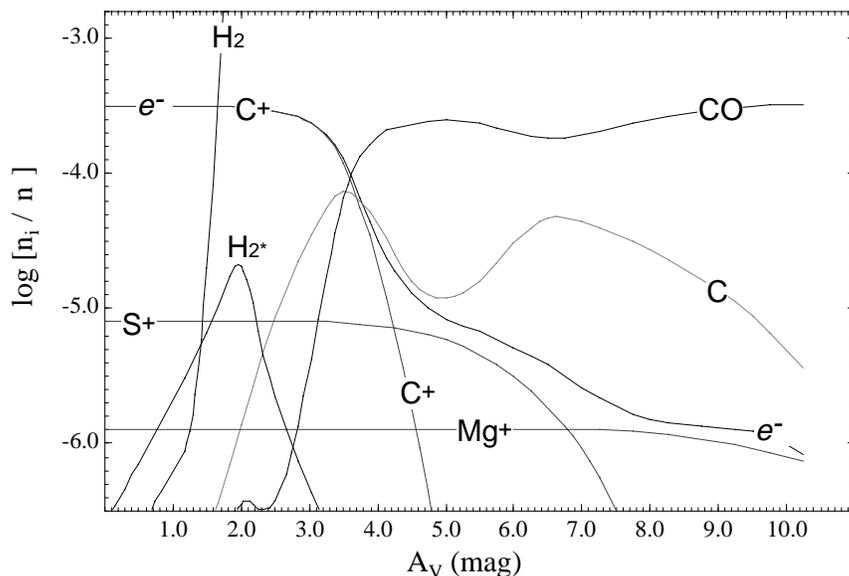
Fig. 13.1. The structure of a photodissociation region in the H, C, H_2 and CO ions. Note that most of the electrons come from the singly ionised atomic species. The electron temperature is of order 1000K out to $A_V \sim 2$ but falls to less than 100K for $A_V > 4$ as CO becomes the dominant coolant. After Tielens & Hollenbach (1985).

recently reviewed by Hollenbach & Tielens (1997, *see* notes). Since the rate of passage of molecular species through the photodissociation zone is slow compared with the various reaction rates, and in particular, the timescale for fomation of molecular hydrogen on grains ($\sim 10^9/n$ yr), a stationary (time-independent) solution of the chemistry can be used instead of the more complex kinetic solution. With the time-independent model, the temperature, ionisation state and molecular composition becomes a unique function of dust extinction optical depth or of hydrogen column density for any assumed set of parameters.

The chemistry of the photodissocation region takes place at a pressure which is equal to the sum of the pressure in the ionised gas and the pressure due to the recoil momentum of the gas streaming off the ionisation front which lies on the inner edge of the photodissociation region facing the exciting stars. This pressure is a factor of order 100 times higher than the pressure of the undisturbed molecular cloud. The photodissociation region is therefore terminated by a shock deep within the molecular cloud (A $_V > 10$ mag). This will usually be a C-type shock.

Thanks to its self-shielding, molecular hydrogen is the last molecular species to be destroyed; at about $A_V \sim 2$. At lower exctinction optical depths

than this, the gas is atomic and is heated by photoelectric emission from dust to a temperature of several thousand degrees K. Such regions are the principal region of emission of the HI 21 cm line, which is the line most used in tracing the dynamical structure of galaxies.

In the region where molecular hydrogen may first form, the gas is still heated about 1000 K. Because here the far-UV pumping rate is high, collisional and radiative excitation of vibrational and rotational levels makes excited molecular hydrogen (H_2^*) an important coolant in the outer boundary of the molecular zone. With increasing optical depth into the cloud the decrease in the far-UV pumping rate leads to a decreasing fraction of the molecular hydrogen in vibrationally excited states.

The vibrational excitation of molecular hydrogen can be sufficient to overcome the activation energy of neutral-neutral reactions, so that chemical reactions such as

$$H_2^* + C \quad \rightarrow CH + H$$
$$H_2^* + C^+ \rightarrow CH^+ + H \tag{13.34}$$

and

$$H_2^* + O \quad \rightarrow OH + H$$
$$H_2^* + OH \rightarrow H_2O + H \tag{13.35}$$

can proceed rapidly in this zone.

Deeper within the molecular cloud, CO can survive, and it rapidly becomes the main coolant, lowering the electron temperature to less than 100K. In this region, the CH_n and the OH_n chain reactions that dominate in dense cosmic-ray ionised clouds are again important. However, photodissociation reactions and dissociative recombination reactions enable additional chemical routes. For example, where there are sufficient electrons present, the OH_3^+ ion produced in the OH_n chain may itself make a dissociative recombination to produce OH or (less likely) H_2O:

$$OH_3^+ + e \rightarrow OH + H_2$$
$$\rightarrow H_2O + H \tag{13.36}$$
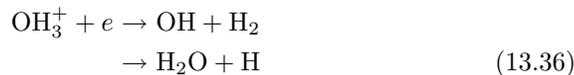
The H_2O and OH produced in this way may then be photodissociated, H_2O to OH and OH to O. This route is countered by the neutral-neutral reactions involving excited molecular hydrogen (13.35). In a similar way, the CH_3^+ ion produced in this chain may also make a dissociative recombination with an electron to produce CH or CH_2:

$$OH_3^+ + e \rightarrow CH + H_2$$
$$\rightarrow CH_2 + H \tag{13.37}$$
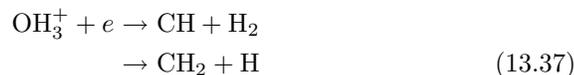
Chemical fractionation of isotopes may be important either in photodissociation regions, or in dense cloud regions. This results from significant differences in the zero-point vibrational energies. For CO the reaction:

$$^{13}C^+ + \ ^{12}CO \rightarrow \ ^{12}C^+ + \ ^{13}CO + \Delta E \qquad (13.38)$$
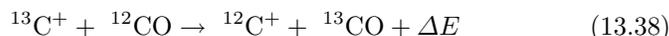
is exothermic, generating an energy $\Delta E/k = 35K$. Although this could, in principle lead to fractionation of CO near the surfaces of clouds, this is supressed at the high temperatures, and is more than compensated for by selective photodissociation by the far-UV field (van Dishoeck & Black, 1988; Keene *et al.* 1998). More important is the fractionation of deuterium in dense clouds (Dalgarno & Lepp, 1984). At low temperatures this proceeds through the reaction:

$$H_3^+ + HD \rightarrow H_2D^+ + H_2 + \Delta E \qquad (13.39)$$
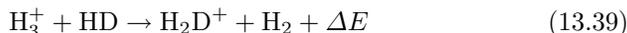
where $\Delta E/k = 200K$, and at higher temperature reactions such as:

$$CH_3^+ + HD \ \rightarrow CH_2D^+ + H_2 + \Delta E \qquad (13.40)$$
$$C_2H_2^+ + HD \rightarrow C_2HD^+ + H_2 + \Delta E \qquad (13.41)$$
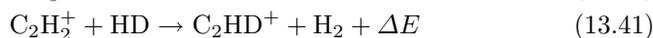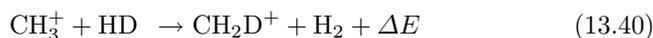
become more important and lead to significant fractionation even at temperatures as high as 70K. The complete fractionation chemistry is described in Millar, Bennett & Herbst (1989).

Despite its very low relative abundance ($\sim 2 \times 10^{-5}$) the HD molecule can, in principle, be detected through its IR fluorescence in photodissociation regions. Although the H_2 molecule can also fluoresce in this waveband, unlike HD it lacks a permanent dipole moment, so it fluoresces with an efficiency a thousand times lower (Sternberg, 1990).

Apart from occurring around hot young stars, photodissociation regions may also be found in the molecular material in the shells of planetary nebulae, and in the hot torii surrounding active galaxies. The active galaxy sources are interesting because in these, the photodissociation regions are illuminated by a very hard and penetrating photon spectrum. As a consequence, the molecular gas is kept warm to greater distance, allowing water vapour to form by the reaction (13.35). The formation of a high H_2O column density which is kept warm by the X-ray photon heating provides ideal conditions for pumping the H_2O megamasers associated with active nuclei (Maloney, 1997); *see* Section (4.3.2).

In X-ray ionised photodissociation regions, the increased ionisation rate accelerates the formation of molecules, but ultimately the X-rays play a more destructive than constructive role as far as molecule formation is concerned (Lepp & Dalgarno, 1996; Sternberg *et al.* 1997). However, the diatomic molecules, particularly the hydrides, are very resistant to destruction by X-ray radiation and are therefore abundant in such zones.

13.2.3 Shock Chemistry

Fast shocks with velocities above ~ 50 km s^{-1} destroy the molecules which enter them. Shocks with velocities above ~ 100 km s^{-1} will fully ionise the gas passing through them. At first sight, therefore, such shocks represent unpromising sites for interstellar chemistry. However, this turns out not to be the case. On the contrary, as shown by Neufeld & Dalgarno (1989), the warm (several thousand degrees K), dense and compressed ($n\,(\mathrm{H_2}) \sim 10^{4-6}$ cm^{-3}) layer of gas that follows the recombination zone in such fast radiative shocks creates ideal conditions to form molecules, since the temperature of the gas is high enough to overcome the activation energies of a number of reactions. In addition, this region is pervaded by a far-UV field generated by the cooling plasma, so many of the chemical reactions seen in photodissociation regions are operative. In particular, rotationally and vibrationally excited molecular hydrogen plays an important role in the chemistry.

Unlike the case of photodissociation regions, in which molecules formed in dense and cool regions find themselves in progressively hotter regions and stronger far-UV radiation fields, the molecules in fast radiative shocks have to be formed in the flow, and move progressively to lower temperature regions with weaker far-UV fields. Since molecular hydrogen is the key to the subsequent chemistry, its formation in the warm (8000-1000K) gas within and following the recombination zone of the shock is essential, since gas phase reactions would proceed to slowly at lower temperature. Under these conditions, molecular hydrogen formation is moderated by the H$^-$ ion formed through the reaction:

$$\mathrm{H} + e \rightarrow \mathrm{H^-} + h\nu \qquad (13.42)$$
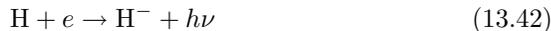
which then permits the formation of molecular hydrogen directly through the *associative detachment* reaction:

$$\mathrm{H^-} + \mathrm{H} \rightarrow \mathrm{H_2} + e \qquad (13.43)$$

the efficiency of this transformation is limited (Dalgarno & McCray, 1972) by the neutralisation of the H$^-$ ion:

$$\mathrm{H^-} + \mathrm{H^+} \rightarrow 2\mathrm{H} \qquad (13.44)$$

Molecular hydrogen is also destroyed at high temperatures by *collisional dissociation*,

$$\mathrm{H_2} + \mathrm{H} \rightarrow 3\mathrm{H} \qquad (13.45)$$

and, in its $v = 2$ state, by the (strongly exothermic) *dissociative charge transfer* reaction with helium:

$$\mathrm{H_2} + \mathrm{He} \rightarrow \mathrm{H} + \mathrm{H^+} + \mathrm{He} \qquad (13.46)$$

Molecular hydrogen may also be transformed to the H$_2^+$ ion through the charge exchange reaction

$$H_2 + H^+ \rightarrow H_2^+ + H \tag{13.47}$$

which then opens the way to the formation of the H_3^+ ion through the reaction given in equation (13.22). Although this may then go on to form the OH^+ or CH^+ ion, the appreciable ionisation fraction in the warm gas means that the hydrogen abstraction reactions such as those given in equations (13.34) and (13.35) will form these ions more rapidly

$$O^+ + H_2^* \rightarrow OH^+ + H$$
$$C^+ + H_2^* \rightarrow CH^+ + H \tag{13.48}$$
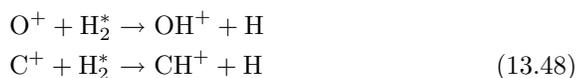
in this respect, and in the activation of the OH_n and CH_n chains, with the electron-induced dissociative recombinations (13.36) and (13.37), the carbon and oxygen chemistry of fast shocks is very similar to that of photodissociation regions.

In, and immediately behind, the recombination region of the shock, atomic cooling dominates (Dalgarno & McCray, 1972), and the temperature falls slowly until at about 4000K. At this point, enough molecular hydrogen has been formed to allow the fractional OH abundance build up to $\sim 10^{-6}$ through the first of the reactions of equation (13.35). Rotational excitation of OH is then the dominant coolant, and the temperature falls rapidly to about 1000K. What happens next depends upon whether the grains have, or have not been, destroyed in the shock; *see* Sections (12.5.1) and (12.5.2). If the shock is faster than about 400 km s^{-1}, then sputtering will have destroyed the grains, the far-UV radiation field will be strong, and the gas will continue to cool with a low fractional abundance of molecular hydrogen. At lower velocities, however, the grains will have been shattered and will present a large surface area to encourage the surface formation of molecular hydrogen. This generates heat of formation, which counters the cooling due to OH rotational excitation to form a plateau at about 400K in the temperature structure, where the formation of the H_2O molecule proceeds rapidly through the hydrogen abstraction reaction (13.35) :

$$H_2^* + OH \rightarrow H_2O + H$$
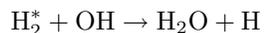
This region of the shock provides high column densities with a low velocity gradient (Elitzur, Hollenbach & McKee, 1989); ideal conditions for pumping the H_2O masers seen around young stars; *see* Section (4.3.2).

Shocks with grain-surface moderated chemistry are capable of building up a richer and more complex set of molecular species, especially complex hydrocarbons such as C_2H_2, $C_2H_3^+$ and C_3H_2.

The shock chemistry in C-shocks, *see* Section (8.4) is somewhat different. Here the degree of ionisation is initially low and in the early part of the shock, nearly all of the oxygen which is not locked into CO is transformed into water by the hydrogen abstraction reactions (13.35). However, as the gas cools, this

water is gradually transformed into OH in reactions with ions, which is in turn dissociated by further reactions with ions.

Silicon is released into the gas phase in such C-shocks through non-thermal sputtering of grains (Pineau de Forets & Flower, 1997), and is rapidly oxidised and transformed into SiO through the neutral-neutral reactions in the warm post-shock gas:

$$Si + O_2 \ \rightarrow SiO + O$$
$$Si + OH \rightarrow SiO + H \tag{13.49}$$

This explains the high column density of this species 10^{13}-10^{14} cm^{-2} which is seen in such shocked molecular outflows (Zhang $et\ al.$ 1995). The SiO subsequently reacts with OH to form SiO$_2$:

$$SiO + OH \rightarrow SiO_2 + H \tag{13.50}$$

which is eventually readsorbed onto the grains in the cool post-shock region.

The column density of SiO in fast J-shocks remains much lower than in C-shocks, because silicon and carbon are largely photoionised in the molecular formation region. As a result of this, much of the silicon forms instead the SiO$^+$ ion, which is rapidly destroyed by dissociative recombination, and what SiO is formed is destroyed by C$^+$ to form the more strongly bound CO molecule:

$$SiO + C^+ \rightarrow CO + Si^+ \tag{13.51}$$

13.2.4 Hot Molecular Cores

Hot molecular cores (HMCs) are the cores of molecular clouds caught just before, or in the early stages of the collapse of the cloud to form a new, frequently massive, star. The most famous examples are the Orion KL region the W51 region and the source Sgr B2(N) in the Galactic Centre region. The observations, physics and astrochemistry have recently been reviewed by Millar (1997) and by Ohishi (1997).

HMCs are seen to be associated with regions of massive star formation, and are very small ($\lesssim 0.1$ pc), dense ($10^6 \lesssim n(H_2) \lesssim 10^8$ cm^{-3}) and optically-thick ($A_V \sim 10^3$ mag), corresponding to hydrogen column densities in excess of 10^{23} cm^{-2}. Their molecular gas is quite warm; 100 - 300K, and they are heated either from the conversion of gravitational potential energy to heat energy, or from shocks. In such dense regions cosmic ray ionisation becomes ineffective as either an ionisation or heating process, and the complex molecules seen in the gas phase have been liberated from the grain mantles by thermal processes in a dense, essentially un-ionised gas. Molecules observed in HMC regions include long-chain species such as HC$_7$N and saturated organic molecules such as H$_2$O NH$_3$, and (dust) "grain" alcohols such as CH$_3$OH and CH$_3$CH$_2$OH.

In the temperature regime appropriate to the hot molecular cores, chemical fractionation effects (*see* Section (13.2.2)) should be small (Millar, Bennett & Herbst, 1989). However, Gensheimer *et al.* (1996) observe the HDO/H_2O ratio to be enhanced in these regions by factors of ten or more over the cosmic ratio of D/H of 1.6×10^{-5} ; *see* Section (4.1.2). This is direct observational evidence that the HDO and H_2O were originally formed as ices on cold grain surfaces in the dense pre-collapse cloud ($T \sim 10$ K). Other ices formed at this time included C_2H_2, CH_4 and NH_3.

After evaporation from the grain surfaces these stable molecules can survive unchanged for at least 10^4 yr. Eventually some of these species react in neutral-neutral reactions to form (possibly even more complex) daughter products. The abundance of these increases with time according to the chemical models described by Millar (1997). For example, the vibrationally-excited HC_3N seen in many sources is formed *via* the reaction:

$$C_2H_2 + CN \rightarrow HC_3N + H$$

Notes on Chapter 13

- The book by Emma L.O. Bakes, *The Astrochemical Evolution of the Interstellar Medium*, (Twin Press:Vledder), ISBN 90-5598-002-1 (hardback) or ISBN 90-5598-003-X (paper) provides an excellent introduction to the field.
- For recent reviews in astrochemistry, see: Dalgarno, A. 1987, in NATO ASI Series, *Physical Processes in Interstellar Clouds,* eds. G.E. Morfill & M. Scholer, (Reidel:Dordrecht), ISBN 90-277-2563-2, p219 and also Genzel, R. 1992 in *The Galactic Interstellar Medium* the lecture notes of the 21st. Saas-Fee Advanced Course, eds. D. Pfenniger & P. Bartholdi, (Springer:Berlin), ISBN 0-3711-55805-5.
- Th key source of rate coefficient data for astrochemical reactions is that maintained by UMIST (Millar, T.J., Farquhar, P.R.A. & Willacy, K. 1997, A&ASuppl.Ser., 121, 139). This may be accessed electronically via: `http://saturn.phy.umist.ac.uk/`. Molecular excitation rates for some intersting species are to be found at `http://www.giss.nasa.gov/data/mcrates/`.
- Photodissociation regions were recently reviewed by Hollenbach, D.J. & Tielens, A.G.G.M. 1997, Ann. Rev. A&A, 35, 179.
- The observations physics and astrochemistry of hot molecular cores has been reviewed by Ohishi, M and by Millar, T.J. 1997, in IAU Symp. #178, *Molecules in Astrophysics: Probes & Processes*, ed. E.F. van Dishoeck, (Kluwer:Dordrecht), ISBN 0-7923-4538-8 (hardback) or ISBN 0-7923-4538-X (paper).

Exercises

14. Thermal Phases of Diffuse Matter

"Double, double toil and trouble,
Fire burn and cauldron bubble"
— Shakespeare (Macbeth, Act 4).

The interstellar medium is a restless place. Heated by supernova shocks and violent stellar winds; cooling and collapsing into dense molecular clouds which form new stars, new winds, and new supernovae in their turn, it is roiled, churned, and compressed to produce a ramified skein of cooler atomic or molecular gas embedded in hotter ionised phases. The spatial structure these various thermal phases is driven by the interstellar gas seeking to attain a pressure balance and a dynamic balance between heating and cooling. Many of these phases are in a stochastic pressure balance with one another.

For our Galaxy the average thermal pressure in the local disk is $P/k = nT \sim 2000 - 6000$ cm^{-3}K (Jenkins, Jura & Loewenstein, 1983). However, the pressure due to the magnetic field is probably of the same order of magnitude, since turbulence will tend to produce equipartition between the magnetic and thermal pressures. The interstellar gas may find itself in any one of five thermal phases. From coolest to hottest these are:

- *The Molecular Medium (MM).* The temperature of this phase is typically only about 20K. Although this phase occupies only a very small fraction of the total volume, of order 1%, it nevertheless accounts for a substantial fraction ($\sim$ 30 - 60%) of the total mass contained in the Galactic ISM, since the density of molecular clouds exceeds 10^3 cm^{-3}. As we have seen in the previous chapter, $A_{\rm V}$ must exceed about 10 in the vicinity of hot young stars to provide full self-shielding against photodissociation. Many of these molecular clouds are bound by their self-gravity, and their dense cores may be in gravitational collapse, forming new generations of stars.
- *The Cold Neutral Medium (CNM).* This material is distributed in dense sheets or filaments occupying 1-4% of the total volume, and is most readily studied by absorption line observations. This material, with a temperature of about 100K is not gravitationally bound, but is in approximate pressure balance with the other components of the ISM. The density of the CNM clouds is therefore 20 - 60 cm^{-3}.
- *The Warm Neutral Medium (WNM).* This phase is traced by H I 21cm emission, and occupies a substantial fraction of the total volume ($\sim$ 30 -

60%) of the gas in the disk of our own, and of other galaxies. It is located mainly in photodissociation regions. In regions distant from young stars, it has $T \sim 6000K$ and $n \sim 0.3$ cm^{-3}, but in photodissociation regions surrounding H II regions the density may be much higher. The heating requirements of this phase are considerable, and it provides a major radiative energy sink for the ISM.

- *The Warm Ionised Medium (WIM).* This phase is most clearly associated with H II regions where it is photoionised by hot young stars. These regions occupy only about 2-4% of the total volume, they have a density greater than 1.0 cm^{-3}, and a temperature is 6000 - 12000K, depending on the gas-phase abundances and on the temperatures of the exciting stars. A WIM has also been detected in regions distant from star-forming regions in both our own Galaxy (Reynolds, 1980, 84) and in external galaxies. This material, sometimes called the *Reynolds Layer* in our galaxy, has $T \sim 6000K$ and $n \sim 0.3$ cm^{-3}, and occupies at least 15% of the total volume. Like the WNM, the heating requirements for this material are considerable, but the nature of the heating source remains uncertain. Near the midplane of disk galaxies, photoionisation seems to dominate, but shock heating or suprathermal particle heating probably become more important as we move out of the galactic plane.

- *The Hot Ionised Medium (HIM).* This phase is heated by strong shocks resulting from supernova explosions or violent stellar winds, and has $T > 10^6K$ and $n < 10^{-2}$ cm^{-3}. As a consequence of its long cooling time, it can come to occupy a large fraction of the ISM. Near the midplane of galaxies, it has a relatively small filling factor, and is largely confined within giant shell H II regions. Here, it becomes detectable in its diffuse soft X-ray emission, thanks to its relatively high pressure in these regions. Above and below the galactic plane, its buoyancy encourages it to bubble up or flow out through galactic chimneys in the denser phase of the ISM, so that it comes to occupy a much larger fraction of the available volume. Here it may be additionally heated by Type I supernova explosions, which originate in an old population of stars with a large scale height. In the halo, the HIM may be detected in absorption against distant stars or galaxies in the highly-ionised species such as N V or O III in regions, provided that it can cool. Eventually this phase merges into the hot halo, with a scale height of 2 - 5 kpc in our galaxy; *see* Section (7.6.3).

In this chapter, we will investigate how the various heating and cooling processes set up and maintain the phase structure of the interstellar gas in galaxies.

14.1 Phase Stability

Within any unit mass of diffuse matter, there will be both a flux of heat in from external sources such as stars, shocks and cosmic rays while at the same time, this same parcel of gas will be losing energy by the various radiative processes described in this book. Both the heat input and the heat output, as we have seen, depends upon the (hydrogen) number density, the temperature, and the state of ionisation, x_i. At densities where collisional de-excitation can be neglected, the loss rate varies primarily on the square of the density, so that the heat loss per unit volume can be written $n^2 \Lambda (T, x_i)$. However, most heating processes (photoionisation, cosmic rays etc.) depend on the first power of the density, so that the heat gain can be written $n \Gamma (T, x_i)$. Thus, the condition for thermal balance is that the net cooling rate $\dot{Q}$ is identically equal to zero:

$$\dot{Q} = n^2 \Lambda - n\Gamma \equiv 0 \tag{14.1}$$

If we consider the P, V diagram for a unit mass of material, then in general, the thermal equilibria represented by this equation will form a line separating regions in which $n\Lambda > \Gamma$ from those in which $n\Lambda < \Gamma$.

Not all these equilibria will be thermally stable, *see* Section (7.5.1). In particular, the gas will be unstable to isobaric (constant pressure) perturbations if the Field (1965) criterion, equation (7.42), is not satisfied, so that:

$$\left[\frac{\partial \dot{Q}}{\partial T} \right]_P < 0 \tag{14.2}$$

while the gas will be thermally unstable to isochoric (constant density) perturbations if the Parker (1963) criterion equation (7.41), is not satisfied, so that:

$$\left[\frac{\partial \dot{Q}}{\partial T} \right]_V < 0 \tag{14.3}$$

The various classes of thermal equilibria which may exist in interstellar gas are shown on the P, V plane in Figure (14.01). Here the line of thermal balance for the plasma is shown in bold. Below this curve and to the left, the cooling rate is less than the heating rate, and above the line and to the right cooling dominates. A number of relevant isotherms are also shown. In panel (a), the gas is everywhere thermally stable. In the region between points A and B, a cold phase with temperature T_1 exists. Below a minimum pressure, $P_{\min}$ the gas becomes warmer as the pressure or density decreases.

In panel (b), there are two thermally stable phases with temperature T_1 in the cold phase and T_2 in the warm. These two phases can co-exist with each other in the pressure range between $P_{\min}$ and $P_{\max}$. In the region between the points C and D, the gas is thermally unstable to isobaric perturbations, since equation (14.2) is satisfied for such perturbations. How this leads to
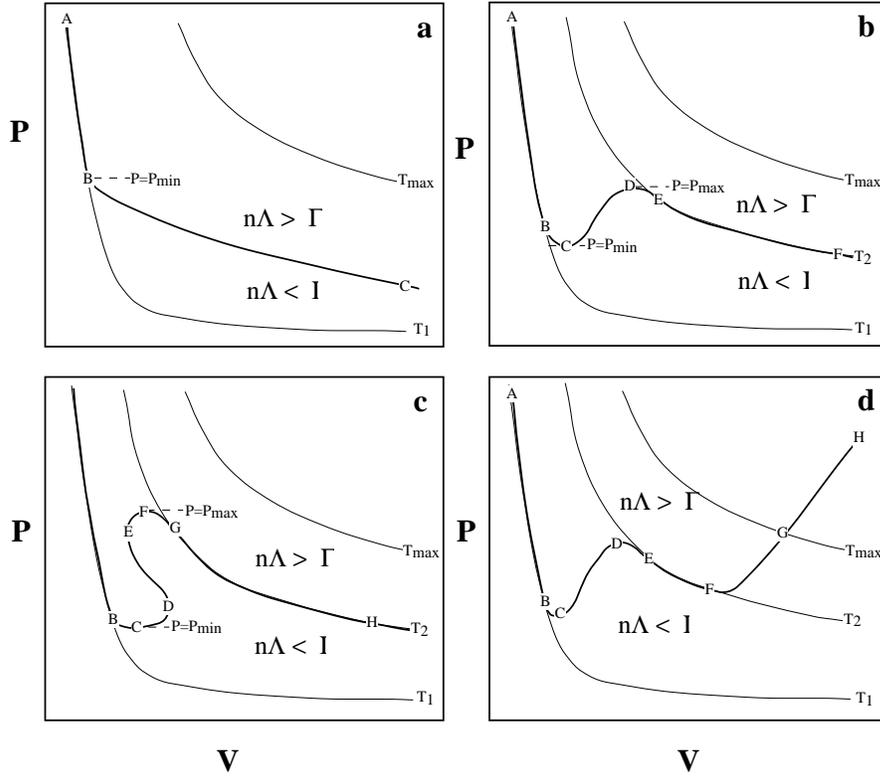
Fig. 14.1. The different types of thermal equilibria which may exist in the inter-stellar medium. In (a) there is one stable phase, which has constant temperature for $P > P_{\min}$. In (b), there are two stable phases separated by a region of instability. In (c) there are also two stable phases, but the region in between can be both isobarically and isochorically unstable. In (d) we have a case like (b), except that there is a further unstable phase at low density. For this phase, when $T > T_{\max}$, the temperature above which the cooling timescale is long in comparison to the evolution timescale, the hot phase can exist in its own right.

thermal instability is very easy to understand. Any small perturbation at constant pressure towards the left will put the gas in a region where the cooling exceeds the heating, and so consequently the gas will continue to cool until finally it comes into equilibrium on the curve A-B. On the other hand, if perturbed towards the right, it will fall into a region where heating exceeds cooling, and so will continue to warm up and move towards the right until it comes into equilibrium on the segment E-F.

In panel (c), there are also two thermally stable phases with temperature T_1 in the cold phase and T_2 in the warm. These two phases can co-exist with each other in the pressure range between $P_{\min}$ and $P_{\max}$. However, in

this case the gas is not only thermally unstable to isobaric perturbations in the region between C and F, but, since equation (14.3) is satisfied, it is also unstable to isochoric perturbations (which are vertical perturbations on this diagram) between the points marked D and E.

Finally, in panel (d) we have two stable phases in the region A-B and E-F, and two regions of instability to isobaric perturbations in segments C-D and F-G-E. In the segment F-G-E the density progressively decreases as the temperature increases. In this case the cooling timescale becomes long at high temperature. Above a certain temperature, say, $T_{\max}$, the cooling timescale becomes longer than any timescale of interest. This may be either the dynamical evolution timescale, the re-heating timescale, or even the age of the universe. Thus, the region between G and H, although formally unstable, is sufficiently long-lived that it could be regarded as defining a third hot phase. If this phase can exist in the pressure regime between $P_{\min}$ and $P_{\max}$ then it is possible to have a three-phase equilibrium consisting of cold, warm, and hot phases.

Under the appropriate conditions, all of these types of phase diagram may be found in real interstellar plasmas. The first two-phase equilibrium to be discussed was by Field, Goldsmith & Habing (1969) for neutral gas heated by cosmic rays. Since that time many other examples have been discussed in the literature, and these are summarised in the following sections.

14.2 Thermal Phases of Galactic Interstellar Gas

14.2.1 Giant Molecular Clouds

The molecular medium of the galaxy is largely concentrated into the giant molecular clouds (GMCs), studied principally through their CO emission. In our solar neighbourhood these typically have masses $\sim 10^5 M_\odot$, are about 45 pc in diameter, and have a surface density in molecular hydrogen of about 10^{22} cm^{-2}. These entities are self-gravitating, so that they are coupled to the stochastic pressure in the interstellar medium only at their surfaces, as a boundary condition. Although individual molecular cloud masses may range from $\sim 10^4 M_\odot$ up to about $\sim 2 \times 10^6 M_\odot$ (Solomon *et al.* 1987) they have remarkable constant surface densities. In the Galactic disk, the molecular clouds have a relatively small scale thickness, about 120 pc (full width half maximum; Stark, 1984; Bronfman *et al.* 1988), and they are the birthplace of both the massive and the low-mass stars. Indeed, star formation is found to be so ubiquitous within them, that star-formation could be taken as one of the defining characteristics of a GMC (Dame *et al.* 1986; Maddelena & Thaddeus, 1986).

Molecular clouds contain very large point-to-point density contrasts and appreciable internal turbulence. The observed relationship between the one

dimensional velocity width, ΔV, (FWHM) and the cloud radius, R (Dame *et al.* 1986; Solomon *et al.* 1987; Scoville *et al.* 1987) can be represented by:

$$\left[\frac{\Delta V}{\text{kms}^{-1}}\right] = 0.7 \pm 0.4 \left[\frac{R}{\text{pc}}\right]^{0.5} \tag{14.4}$$

It is generally accepted that the GMCs are self-gravitating entities, and that it is this turbulence that is preventing them from undergoing rapid gravitational collapse. The origin of the turbulence in GMCs is probably to be found in the outflows from young, low mass stars which give rise to the bipolar molecular flows and to the Herbig-Haro objects. These outflows produce magnetic-field limited compression and vorticity which is dissipated in a self-similar way. In the Orion south molecular cloud (Fukui *et al.* 1986), and the Mon OB1 GMC (Lada, 1988) the turbulent energy input supplied by the CO outflows is sufficient to balance the cloud provided that the turbulent dissipation timescale is an order of magnitude longer than the free-fall timescale. These processes are reviewed by Falgarone (1995). If the clouds are in virial equilibrium (where the turbulence supports the cloud against its self-gravity), then:

$$\Delta V^2 = \alpha \frac{GM}{R} \tag{14.5}$$

where M is the mass of the GMC, and α is a constant depending upon the internal mass distribution of the cloud. Equation (14.4) then implies that $M \propto R^2$, and that the mean surface density of all the GMCs are similar. This must be telling us something about how the GMCs regulate themselves. Chièze (1987) argues that this scaling relation is exactly what would be expected if GMCs are on the edge of gravitational instability in a constant pressure environment, with sub-condensations forming a virialised N-body system. In this case,

$$\left[\frac{M}{M_\odot}\right] \sim 160 \left[\frac{P/k}{5000\text{cm}^{-3}\text{K}}\right]^{1/2} \left[\frac{R}{\text{pc}}\right]^2 \tag{14.6}$$

where P is the pressure external to the cloud. An alternative viewpoint has been advocated by McKee (1989), who notes that the mean surface density is just what is needed to fully absorb the UV stellar radiation field, and who therefore suggests that the star formation in GMCs is essentially photo-regulated.

The evidence for large density contrasts comes from multiline CO studies, observations of the ^{13}CO (1-0) line, which tends to be optically thin, and analysis of IRAS observations. These suggest (*i.e.* Falgarone & Puget, 1986; Devereux *et al.* 1994) that of order, or even more than, 90% of the mass of the GMCs resides in the clumps, and that these clumps have H_2 densities $\sim 10^3$ cm^{-3}; compared with the inter-clump density of only $\sim 3 - 10$ cm^{-3}. Such a large density contrast suggests that individual GMCs may be themselves regarded as constituting a two-phase medium (McKee, 1995; *see* notes on

this chapter). Here, the inter-clump medium is composed mainly of warm atomic gas heated by the stellar radiation field. The clumps are moving in this medium, and are gravitationally bound together as an ensemble (although the self-gravity of each clump may not be very important in setting the physical conditions within it). Certainly in this case the thermal pressure in the interclump medium must increase towards the centre to provide the required hydrostatic support against gravity. In a cloud with a density profile $\rho \propto r^{-1}$ the relationship between the central pressure, P_c, and the pressure at the surface, P_o; the local thermal pressure of the substrate interstellar medium is (McKee, 1989):

$$\frac{P_c}{P_o} = \exp\left[\frac{\phi}{c_1^2}\right] \tag{14.7}$$

where ϕ is the depth of the central gravitational potential well, and c_1 is the isothermal sound speed in the interclump medium. For most GMCs, the pressure at the cloud centre is at most a factor two higher than at the surface. The pressure in the clumps would then match the local pressure in the interclump medium, provided that the clumps are not dominated by their own self-gravity, which implies a clump density of $\sim 10^3$ cm^{-3} for a clump temperature of 10K; figures which are entirely compatible with observation.

14.2.2 The Atomic ISM

Advanced Section

The atomic phases of the interstellar medium of our Galaxy comprise, in the terminology originated by McKee and Ostriker (1977) the cold neutral medium (CNM) and the warm neutral medium (WNM). Much of the mass resides in the CNM, which has a density of ~ 20 cm^{-3}, a spin temperature of ~ 100 K, and is seen in H I absorption. This phase has only a small filling factor. The WNM is observed in H I emission and has a density of $\sim 0.3 - 1.0$ cm^{-3}, a spin temperature of ~ 6000 K and fills a substantial fraction of the available volume. These two phases exist in stochastic pressure balance with one another, which, as we have seen from Section (14.1) is a necessary condition for the existence of a two-phase medium. A theoretical understanding of these phases was given by Field, Goldsmith & Habing (1969), following on the classical work of Field (1965) on thermal instability. According to them, both phases are primarily heated by cosmic rays, a phase diagram similar to Figure (14.1.b) applies, and the pressure regime lies in the range $P_{\min} > P > P_{\max}$ so that a two-phase equilibrium is possible. In practice, the ionisation rate they assumed, $\zeta = 4 \times 10^{-16}$ s^{-1}, is considerably higher than what is acceptable today, so the source of the heating is probably not soft X-rays. Nonetheless, regardless of the source of the heat input, the conclusions they reached about the phase structure are secure.

The energy requirements for heating the atomic ISM are considerable. The primary coolant of the CNM is the [C II] $^2P^0_{3/2} - {}^2P^0_{1/2}$ transition at

157.74μm. Wright *et al.* (1991) estimate that the luminosity of the galaxy in this line alone is 2×10^{41} erg s^{-1}, or about $10^8 L_\odot$. This is equal to about 10% of the kinetic energy deposited into the interstellar medium by all the supernova explosions in the galaxy. Only three energy sources can provide sufficient energy to heat the CNM and WNM; supernova explosions, starlight or viscous tapping of the rotational energy of the galactic interstellar gas.

Supernova explosions cannot directly heat these phases, since much of the energy deposited by a supernova into its local ISM is frittered away in adiabatic losses and radiative processes. However, the energy given out in the form of soft X-rays and cosmic rays is sufficiently penetrating to be available to heat the neutral phases of the ISM. The soft X-rays have an energy density of only 10^{-5} eV cm^{-3}, which is totally insufficient to provide the required heating. They can therefore be safely ignored in their effect on the phase structure, but they are important in maintaining the ionisation, since each X-ray absorbed may produce many secondary ionisations; *see* Section (5.3.2).

Cosmic rays are produced by Fermi acceleration is supernova shocks, and may carry away a substantial fraction of the shock energy. In the interstellar medium, these have a high energy density, about 1.0 eV cm^{-3}. However, much of this energy density is in the form of energetic cosmic rays which are very penetrating, and which are therefore ineffective in heating the neutral phase. On the other hand the low energy cosmic rays are effective in heating, but have a short stopping length, and their diffusion is much more dependent on the magnetic field configuration. Integrating over the energy spectrum, cosmic rays produce an ionisation rate of $\zeta \sim 2 \times 10^{-17}$ s^{-1} (van Dishoeck & Black, 1986), which is far less than was assumed by Field, Goldsmith & Habing (1969). Such an ionisation rate is insufficient to account for the energy budget of the neutral phases.

The interstellar turbulence and the magnetic field, which is ultimately derived from this turbulence, can provide sufficient heating, in principle. The turbulent velocity dispersion of the H I layer in the galaxy (about 12 km s^{-1}) is equivalent to an energy density of about 4 eV cm^{-3}, and the magnetic energy density is about 1.0 eV cm^{-3}, which is about equal to the thermal energy density in the WNM. If the turbulence generated by stellar winds and supernova explosions can be effectively coupled to the WNM, then here we have an adequate heating source. However, this will only work if the filling factor of the hot medium is large as advocated in the McKee & Ostriker (1977) paper. Tapping the interstellar magnetic field energy is more difficult. Two alternatives have been suggested, magnetic reconnection (Raymond 1992), and hydromagnetic wave heating (Ferriere, Zweibel & Shull, 1988).

By far the most promising heat source is starlight. The local energy density is also about 1.0 eV cm^{-3}, and a considerable fraction of the energy in starlight is absorbed by dust located in the atomic phase. In the solar neighbourhood Mathis, Mezger & Panagia (1983) estimate that the stellar radiation field carries an energy flux of 2.2×10^{-2} erg cm^{-2} s^{-1} or $54 L_\odot$ pc^{-2}.

The proportion of this that is absorbed, mostly in the 912 - 3600Å waveband, directly heats the dust which re-emits in the IR. From IRAS data, Boulanger & Perault (1988) estimate that dust in atomic gas re-emits $9L_\odot$ pc^{-2}, so that at least 17% of the stellar radiation field is being absorbed in this component. The main heating source for the atomic gas is by photoelectric emission from small grains; see Section (12.4), since the large grains are relatively ineffective in this regard (Wolfire *et al.* 1995).

(Ralph: INSERT your discussion of the phase equilibrium (your poster paper with Mike Shull) here!)

14.2.3 The Warm Ionised Phases

The ionised gas in disk galaxies exists in two locations, within obvious H II regions and in a more diffuse and extended low ionisation layer, which nevertheless shows a rough spatial correlation with the H II regions, and which provides a significant fraction, 20-40% of the total galactic Hα luminosity.

Clearly the H II regions themselves must be photoionised by their hot central star. For these, the balance of heating and cooling is as shown in Figure (9.5), from which it is clear that only a single equilibrium temperature is possible. H II regions cannot therefore display a multi-phase equilibrium, although for these more complex *acoustic* or *thermoreactive* instabilities may exist (*see* Ferrara & Corbelli, 1995).

The nature of the more extensive WIM remains more controversial. The energy it requires to maintain it is about equal to all of the mechanical energy put into the ISM by supernovae or stellar winds, about a third of the ionising photon energy put out by young OB stars, or a few percent of the energy carried by starlight. The size of the measured [N II]/Hα ratio measured at high z-heights in external edge-on galaxies is difficult to account for with normal photoionisation models. A mechanical source of energy or energy in some other form is probably needed to account for the excitation of this material. For the rest, photoionisation by OB stars is the most likely energy source.

Recent observations of the DIM in both normal and starburst galaxies by Wang *et al.* (1998) have revealed a striking correlation between the excitation, as measured by the [S II]$\lambda\lambda$6717+6731Å/Hα ratio and the local Hα surface brightness relative to the mean value in the galaxy, $\Sigma_{H\alpha}/\langle\Sigma_{H\alpha}\rangle$. They show that this relation is a natural outcome of a model in which the DIM is photoionised gas that has a mean pressure P that is proportional to the mean star formation per unit area, $d\Sigma_*/dt$. Such a scaling may either arise either because the massive stars pressurise the ISM directly, or because feedback processes in the galactic disk lead to the star formation rate being determined, or limited by the mean gas pressure. This point will be discussed in more detail below.

A true multi-phase equilibrium can exist in the dense gas found in the vicinity of energetic active galactic nuclei where Compton heating is impor-

tant. Here the radiation pressure carried by the ionising radiation field at radius r from the central source, $P_{rad} = L/4\pi r^2 c$ can become as large as, or even larger than, the gas pressure. If $F(r)$ is the local flux in the radiation field (erg cm^{-2} s^{-1}), then it is convenient to define the local ionisation parameter in terms of this pressure ratio; see equation (9.8):

$$\Xi = \left(\frac{P_{rad}}{P_{gas}} \right) = \frac{F(r)}{cknT} = \frac{L}{4\pi r^2 cknT} \tag{14.8}$$

where n is the total particle (electron + ion) density.

The energy balance can be split up into terms of the energy loss due to collisional processes, the energy gain from photoionisation heating, the energy loss due to the inverse Compton process, and the energy gain due to Compton heating, respectively:

$$\dot{q}_{net} = \Lambda - \frac{\Gamma}{n} + \Lambda_C - \frac{\Gamma_C}{n} \tag{14.9}$$

Considering only the Compton heating and cooling, the net cooling rate in the nonrelativistic scattering case is:

$$\dot{q}_C = \Lambda_C - \frac{\Gamma_C}{n} = \frac{\sigma_T F(r) n_e}{m_e^2 c^2 n^2} [4kT - \langle \varepsilon \rangle] \tag{14.10}$$

where σ_T is the Thompson scattering cross section, and the mean photon energy $\langle \varepsilon \rangle$ is:

$$\langle \varepsilon \rangle = \frac{1}{F(r)} \int F(r, E) E dE \tag{14.11}$$

When Ξ is very high, the Compton terms dominate the thermal balance, and the plasma comes into equilibrium at the *Compton temperature*, T_C ;

$$T_C = \frac{\langle \varepsilon \rangle}{4k} \tag{14.12}$$

The exact value of the Compton temperature depends on the shape and maximum energy of the ionising spectrum, but is somewhere between 3×10^6 K and 10^8 K.

At low ionisation parameter, Ξ, the plasma achieves a photoionisation balance such as discussed in Chapter (9), with the plasma at a temperature T_P at which the photoionisation heating and the radiative cooling balance:

$$\Lambda(T_P) = \frac{\Gamma(T_P)}{n} \tag{14.13}$$

Now, let us see how a parcel of gas in a fixed radiation field $S(r)$ can achieve a multi-phase equilibrium under these conditions. At very high pressure, the ionisation parameter Ξ may be so low that the gas is mostly atomic.

Nonetheless, the X-ray heating may be great enough to maintain the temperature near 10^4 K. As the pressure decreases, the gas becomes fully ionised. In the approximate range $-4 < \log \Xi < 1$, the ionisation state increases as the ionisation parameter increases. Since the heat input through photoionisations increases, the temperature also increases so that the cooling rate can match the heating rate. This state of affairs cannot continue indifinately, since eventually the cooling rate will decrease as the main coolants are depleted from the plasma; *c.f.* Figure (7.1). Thus when the temperature is of order 10^5 K, at a pressure $P_{\min}$, and ionisation parameter $\Xi_{\max}$, heating overwhelms cooling, a stable equilibrium is no longer possible, and the temperature runs away up to the Compton temperature.

At low pressure, the ionisation parameter is very high, so that the gas is at its Compton temperature. As the pressure increases, free-free (thermal Bremsstrahlung) cooling becomes significant, and the equilibrium temperature decreases somewhat. Eventually, at temperatures of order 10^7 K, X-ray line and continuum cooling kick in. Eventually, the increase in cooling rate at the temperature decreases is sufficient to render the plasma unstable to isobaric perturbations, and, provided that the cooling timescale is shorter than the dynamical evolution timescale, the gas cools catastrophically back down to the photoionisation equilibrium temperature, at a pressure $P_{\max}$, and ionisation parameter $\Xi_{\min}$, of order $0 - 10$.

It is clear therefore, that a phase diagram similar to Figure (14.1b) or (14.1c) applies, and that between $\Xi_{\max}$ and $\Xi_{\min}$, a stable two-phase medium can exist, with clouds in photoionisation equilibrium between 10^4 and 10^5 K, embedded in an intercloud medium at its Compton temperature at 10^7 to 10^8 K. A discussion of the trajectories of such a plasma on the $P - V$ plane is given by Begelman & McKee (1990). What is not yet certain however, is whether these physical conditions can be actually realised by gas in the vicinity of AGN.

14.2.4 The Hot Ionised Component

Based solely on cooling arguments, the hot phase of the interstellar medium cannot be thermally stable. It derives its existence from the fact that, because its density is so low, its cooling timescale is longer than the timescale over which it is re-heated by supernova shocks. The sun sits in such a hot phase, in a bubble of plasma with a temperature of order 10^6 K, and a hydrogen density $\sim 5 \times 10^{-3}$ cm^{-3}. According to the McKee & Ostriker (1977) model, supernova remnants preferentially evolve in this low density phase because there, the low density gas leads to a high shock velocity. The probability per unit time of a shock passing through a given parcel of gas, $P(t)$ is:

$$P(t) = \frac{1}{\tau_{SNR}} \exp\left[-\frac{t}{\tau_{SNR}}\right] \tag{14.14}$$

where τ_{SNR} is the mean interval between supernova remnant shocks. This is about 5×10^5 yr, locally. The hot medium sweeps over clouds in the cooler phases, shredding them, and heating the remnants by thermal conduction. To a first approximation, therefore, in the evaporative limit the equilibrium temperature of the hot phase is determined by the energy balance between the input shock energy and the conductive losses. Cloud shredding, ablation and thermal conduction provide the means of transporting matter from to cooler to the hotter phase, while radiative cooling in locally overdense regions allows mass transport in the opposite sense.

Thermal conduction serves to stabilise the hot medium against thermal instability. This can easily be seen by considering the evaporative cooling function, Λ_{ev}, expressed in terms of the conductivity, κ, and the cloud area per unit volume A_C :

$$\Lambda_{ev}\left(T\right) = \frac{8\pi}{5} A_C \left(\frac{\kappa T}{n^2}\right) \qquad (14.15)$$

Because the conductivity is such a steeply rising function of temperature; $T^{5/2}$, *see* equation (7.21), the effective cooling rate provided by the thermal conduction thermally stabilises the plasma.

Because the sound speed of the hot medium is so great ($> 100\ \mathrm{km\ s^{-1}}$), the hot medium communicates the background pressure to the other phases of the galactic medium. This communication may occur via the halo as local regions of overpressure are vented through galactic chimneys into the extended hot halo of the galaxy.

The key parameter is the so-called *porosity*, which is the filling factor of this medium. If this is much less than unity, the regions of hot plama remain mostly isolated from one another, like the holes in a Swiss cheese. However, when the filling factor exceeds about 0.5, the hot bubbles become connected, and the cooler phases are compressed into a ramified network of blobs and filaments. This appears the case which most closely corresponds to what we see in our solar neighbourhood, or in the Magellanic Clouds. Finally, as the filling factor of the hot phase approaches unity, the cooler phases are confined to a set of isolated clouds. This is the situation which most likely applies in our hot galactic halo.

14.3 Feedback & Mass Exchange

14.3.1 Shells, Supershells & Interstellar Froth

In earlier sections, we have considered the evolution of bubbles of hot plasma formed in the interstellar medium by supernova explosions or mass-loss from massive stars. In regions of OB associations, these bubbles collide with one another, run together and merge to form larger structures. Indeed, observations in the EUV and absorption line observations have revealed that our

Sun itself is sitting in one such bubble, the Local Hot Bubble (LHB) a cavity with a redius of order 100pc, filled with hot tenuous plasma at a temperature of order 10^6 K, and a hydrogen density $\sim 5 \times 10^{-3}$ cm^{-3}. Egger (1995) further suggests that the so-called North Polar Spur, a looped enhancement of radio nonthermal and X-rays extending to high galactic latitudes, and which is part of the Galactic Loop I represents the region of shock collision between LHB and the bubble blown by the Scorpius-Centaurus OB association.

Here we modify somewhat the nomenclature introduced by Meaburn (1980). Let us first define H I or H II *giant shells* as regions having a radius less than the scale height of the H I gas, which are therefore confined within the main H I layer in the LMC. In larger regions than this, the shape of the bubble will be very strongly modified by the density gradient, being flattened on the side nearest the galactic plane, and extended out of the plane.

Supergiant shells would then be those regions whose extent above the plane is so much larger than the H I scale height that the hot gas produced within the shell has broken out of the plane and is either draining out or has drained out its hot X-ray emitting gas into the hot halo gas around the LMC. Such regions form *galactic chimneys* or *worm-holes* orientated perpendicular to the disk plane. When it is drained of the overpressure which is driving the expansion of the H I shell, it has reached a momentum-conserving phase, and its expansion velocity in the galactic plane will reduce as more and more disk gas is swept up into the expanding H I shell. Unless it is powered by secondary star formation around its periphery, it will lose its identity either when the expansion velocity falls below the random turbulent velocity of the disk gas, or else when the differential velocity shear due to rotation becomes larger than the expansion velocity.

This provides a simple and physically meaningful distinction between the various shell classes. Let us now attempt to quantify this distinction. For an isothermal disk, the scale height, z is related to the velocity dispersion at the mid-plane, v, and the surface density of matter, Σ_T:

$$z = \frac{\langle v^2 \rangle}{\pi G \Sigma_T} \tag{14.16}$$

In the van der Kruit & Searle (1981a,b; 1982) disk model, the solution to the Poisson equation gives a vertical density profile $\rho(h)$ of the form $\rho(h) = \rho(0)\mathrm{sech}^2(h/z)$. At large height, this distribution is indistinguishable from an exponential. However, in a real galaxy the matter is not isothermal, since we have both gaseous and stellar components which, in general, have quite different velocity dispersions. Since the velocity dispersion of the gas component, v_g, is much smaller than the stellar component, v_*, the scale height of the gaseous component, z_g, is also much less than for the stars, z_*. In the particular case that the disk matter has an exponential distribution, and where the gas can be regarded as a sub-population in the same potential, but with different scale height, then, provided that magnetic pressure sup-

Table 14.1. Classification of Shell Structures

Central Stars	Shell Type	dE/dt (erg s-1)	E (erg)	R (pc)
Individual OB/W-R	Bubble/Ring Nebula	10^{36-37}	10^{50-51}	1-30
OB Association	Giant Shell	10^{37-39}	10^{51-53}	50-180
Multiple OB Assns.	Supergiant Shell	$> 10^{38}$	$> 10^{53}$	>180

port can be neglected, the mid-plane velocity dispersion of the gas is given (exactly) by (Dopita and Ryder 1994):

$$v_g^2 = \frac{2\pi G \Sigma_T z_g^2}{(z_g + z_*)} \tag{14.17}$$

For the Galaxy, the scale height of the stars $z_* \sim 350$ pc and the scale-height of the gas layer, $z_g \sim 150$ pc), while for the LMC, the the numbers are not very different; $z_* \sim 550$ pc and $z_g \sim 190$ pc (Kim *et al.* 1998b). We therefore set the division between shells and supershells at a diameter between 300 and 400 pc. With this definition, the nature of the shells, their energy requirements, and the appropriate shell sizes are listed in Table (14.1).

The LMC is an ideal system in which to study the collective effects of star formation and bubble development, since it is seen nearly face-on, and both the H I and the H II have been surveyed at comparable (~ 15 pc) resolution Kim *et al.* (1998a, b).

The expansion velocity of the H I shells shows a very clear correlation with the radius. For the giant shells, the expansion velocity increases from about 15 km s^{-1} for the smallest up to about 20-35 km s^{-1} showing that these are being accelerated by the energy input from the stars within them. The trend of velocity with radius is not an evolutionary one, since it is quite unlike that predicted by the standard Weaver *et al.* (1977) theory given in Section (8.3.2). In fact, the relationship arises from the intrinsic differences in mechanical energy input and ISM density in the giant shells of different sizes.

On the other hand, the supergiant shells show no trend with radius. The break in the expansion velocity versus radius relation occurs at boundary between the two shell classes at $D = 380$ pc, and is consistent with the supergiant shells having exceeded their blowout diameter which we had estimated on the basis of the scale height of the matter in the disk of the LMC.

A high proportion of the giant and supergiant H I shells are colliding with one another. Many of them are interlocking, especially near the 30 Dor complex, where very active star formation has been occurring simultaneously in many different centres. The fact that the H I velocity structure in both the LMC and the SMC is dominated by a series of interlocking and interacting bubbles shows graphically that these are providing a fundamental control of the phase structure of the interstellar medium in these galaxies. This supports the model advocated by Oey & Clarke (1997), who assumed that shell

expansion is initially adiabatic, but stalls when the pressure in the bubble becomes comparable with the background pressure in the ISM. For a constant birthrate and a mechanical luminosity function $\phi(L) \propto L^{-\beta}$, this gives a predicted size distribution $N(R) \propto R^{1-2\beta}$. The slopes predicted by this theory, using the observed luminosity distribution of the H II regions are in excellent accord with those derived directly from observation for both the SMC (Staveley-Smith *et al.* 1997) and the LMC (Kim *et al.* 1998b).

Where H I shells are associated with H II regions, the H I structure is systematically larger than the H II structure in projection. This is to be expected since the H II region is confined in a thin layer between the ionisation front at the inner boundary of the H I shell and the hot shocked stellar wind material ejected by the central stars. In addition, we find that compact H II regions are preferentially located in, or on the boundaries of, dense H I filaments, as would be expected since star formation will naturally take place in the dense regions.

In general, however, there is only a weak correlation between the H I shells and the ionised gas traced out by the H II regions and H II shells. This is shown graphically in Figure (14.2), where the H I is shown in white tones, and the H II as dark tones. Since we believe that massive star formation is the ultimate causal factor of both the H I shells and supershells, this shows that either the lifetime of the H I shells is typically longer than the lifetime of the OB stars which input the energy required to drive the expansion of the H I shells, or that the pressure falls in the ionised regions sufficiently to lower the emission measure of the H II region below the threshold of observability. This very faint ionised meterial constitutes the WIM in galaxies.

14.3.2 Self-Propagating Star Formation

Advanced Section

In the LMC at least, it is clear that a number of the smaller shells have been formed on the rims of supergiant shells. Such data provides clear observational support for the idea of self-propagating star formation, according to which gravitational instabilities in the swept-up material of the supergiant shell caused fragmentation and a new round of star formation (Dopita *et al.* 1985; McCray and Kafatos 1987; Elmegreen 1998; Efremov & Elmegreen, 1998). The best example of such self-propagating star formation in a supergiant shell is the LMC4 / Constellation III complex. This region in the northern part of the LMC appears as a large loop with associated H II regions in the upper left hand quadrant of Figure (14.02). It is nearly 1.4 kpc across and seems to have been initially triggered some ~ 30 Myr ago. This age is based on the existence of a central cluster of A-type supergiant stars (Efremov & Elmegreen, 1998), and this region seems to have passed through at least two episodes of triggered star-formation up to the present day, which first led to the formation of a great arc of stars $\sim 10-18$ Myr ago, and finally to the set of H II regions and H I giant shells and supernova remnants (with
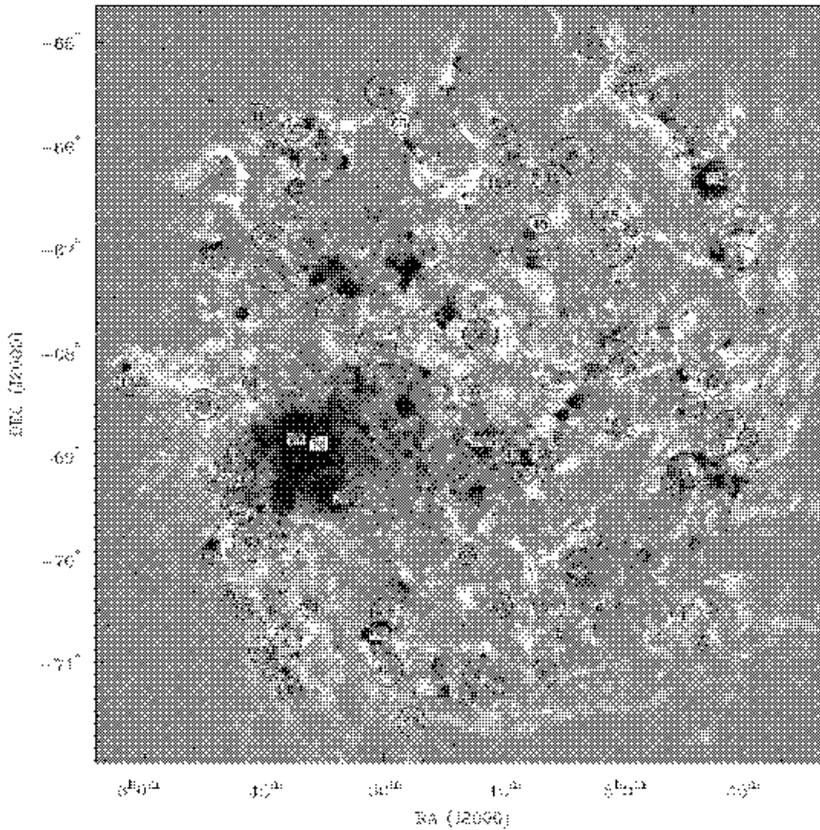
Fig. 14.2. A comparison of the H I map of the LMC (white tones) with the Hα image. Note the very bubbly structure in the H I image, and the manner in which the star forming regions tend to be embedded in dense H I filaments.

ages of $\sim 1 - 10$ Myr) which are currently located on the inner side of the H I supergiant ring.

Although the LMC4 / Constellation III region is the best-known example, the 30 Dor region, the brightest H II complex in Figure (14.2), is also a fine example of the phenomenon. The associated supergiant shell LMC-2, located just to the east of 30 Dor has been studied by Wang & Helfand (1991) and Caulet & Newell (1996).

Self-propagating star formation is a phenomenon which extends down to the scale of giant H II shells, provided they occur in dense enough regions and provided that the local star-formation rate has been high enough. Examples in the LMC include the DEM 34 or N11 region (Walborn & Parker, 1992; Rosado *et al.* 1996) and the DEM 192/ N51D region (Oey & Smedley 1998).

A particularly fine example of a giant shell displaying self-propagating star formation is provided by the complex N44 (Oey & Massey 1995), recently studied by Kim, Chu, Staveley-Smith & Smith (1998). This region contains three OB associations, and a number of single-star H II regions around the rim. The central association LH 47 is clearly older, since it contains a number of red supergiants. There is also a supernova remnant nearby. N44 is filled with diffuse X-ray emission from hot plasma, which has already blown a pair of prominent H I shells. Both the H I and the H II regions are expanding at velocities between 30 and 60 kms^{-1}.

For most giant shells there is a longstanding problem (Oey & Massey, 1995) that the apparent kinetic energy of the expansion is less than is expected by application of the Weaver *et al.* (1977) theory of pressure-driven bubbles. The probable solution to this is that a great deal of the energy of the expansion is stored in the motions of the molecular component which we cannot easily measure.

14.3.3 Self-Regulated Star Formation

In disk galaxies, the most compelling evidence that the star formation rate is self-regulating is the existence of a connection between the local star formation rate in the disk, and the local disk properties. This is usually expressed in terms of a Schmidt (1959) relationship connecting the star formation rate per unit area of disk, Σ_{SFR}, with the surface density of gas, Σ_g:

$$\Sigma_{SFR} = A\Sigma_g^N \qquad (14.18)$$

with the power law index being determined observationally; $0.9 < N < 1.8$.

How are the quantities in this equation measured? The star formation rate has been generally derived from Hα surface photometry. Since this measures the rate of production of ionising UV photons, it is proportional to the luminosity of the OB stars (*i.e.* the number of OB stars times their area), which is in turn proportional to an product of the birthrate of such stars multiplied by their lifetime, weighted in the appropriate way across the initial mass function. A surface brightness of 1.0 $L_\odot$ pc^{-2} in Hα is equivalent to a birthrate of massive stars ($M > 10 M_\odot$) of ~ 3.0 $M_\odot$ pc^{-2} Gyr^{-1}, which corresponds to about 20 $M_\odot$ pc^{-2} Gyr^{-1} when the integral is taken over the whole of the initial mass function (Kennicutt, 1983; Dopita & Ryder 1994). More recently Kennicutt (1998) has used the far-infrared (FIR) luminosities of galaxies, to allow the calibration to be extended to dusty starburst systems in which much of the Hα may be absorbed internally to the galaxy (when $\Sigma_H \gtrsim 50 M_\odot$ pc^{-2}). This FIR calibration relies upon the fact that the IR emission is largely due to the absorption of the UV starlight and the Lyα photons produced in the ionised gas by the surrounding dusty gas. Since both of these depend on the luminosity of the OB stars, then like the Hα luminosity, the FIR luminosity measures the stellar birthrate. A surface luminosity of

1.0 $L_\odot$ pc^{-2} in the FIR is equivalent to a total birthrate of $\sim$5.8 $M_\odot$ pc^{-2} Gyr^{-1} for starbursts lasting at least $10^7 - 10^8$ yr in which the star formation can be regarded as continuous.

The most direct way of measuring the gas is through the H I. However, as we have seen, H I is located in the photodissociation regions, while much of the mass in the ISM may be concentrated in the molecular clouds. The CO may be used as a tracer of this component, but this depends upon the highly uncertain $X-$factor, the conversion ratio between CO surface luminosity and molecular hydrogen column density, which we have already alluded to more than once in this volume.

With these caveats, Kennicutt (1998) has determined the best-fit Schmidt law for both spirals and starburst galaxies. Within the star-forming disk:

$$\left[\frac{\Sigma_{SFR}}{\text{M}_\odot \text{ pc}^{-2} \text{ Gyr}^{-1}} \right] = (0.25 \pm 0.7) \left[\frac{\Sigma_g}{\text{M}_\odot \text{ pc}^{-2}} \right]^{1.4 \pm 0.15} \qquad (14.19)$$

At the edge of the star-forming disk, the star formation rate drops very rapidly, as the gas in the disk falls below the threshold density Σ_{crit} for the self-gravitational instablility to operate (Quirk, 1972; Kennicutt, 1989);

$$\Sigma_{crit} = \alpha \frac{\kappa \langle v \rangle}{\pi G} \qquad (14.20)$$

where κ is the epicylic frequency of the gas in orbit in the galaxy, $\langle v \rangle$ is its mean velocity dispersion, and α is a constant of order unity.

Equation (14.19) implies that, if it were not for the fact that stars recycle their gas back into the interstellar medium, the disks would be depleted of their gas in less than a Hubble time, and in timescales as short as 10^8 years in the case of the luminous starburst galaxies. Clearly then, the starburst condition is not one that can last too long.

What is the physical meaning of the Schmidt Law? The simplest theoretical scenario is one in which the star formation rate is presumed to scale with the growth rate of gravitational perturbations within the disk. In this case, the local star formation rate (per unit volume) will scale as the local gas density divided by the growth timescale, $\rho_{SFR} \propto \rho_g / (G\rho_g)^{-1/2} \propto \rho_g^{3/2}$. The scaling to surface quantities depends upon the local scale-height of the gas layer, but it is plausible that this may produce an N in the right range.

An alternative viewpoint is that all the star formation in the disk is self-propagated, with supernova explosions serving to compress dense clouds and trigger the formation of new stars in their vicinity. This picture is closer to the self-regulation model that the observations of the LMC suggest. Sleath & Alexander (1995) have built a computer simulation in which the molecular cloud mass distribution is taken to be a power-law, and the probability of stimulated star formation is assumed to increase proportionally to the mass. They find that this reproduces a Schmidt Law with index $= 1.65$.

The Schmidt law, expressed in the form given above may however not be unique, or the only physical reasonable form, even if it does serve to provide a good fit to the observations. For example, a number of authors (Larson, 1988; Wyse 1986; Silk 1997 and Elmegreen 1997) have suggested that, rather, the star formation may scale with a power the gas density divided by the local dynamical (orbital or infall) timescale. For a disk galaxy;

$$\Sigma_{SFR} \propto \frac{\Sigma_g^N}{\tau_{dyn}} \propto \Omega \Sigma_g^N \tag{14.21}$$

where Ω is the local angular rotation speed. Both from theoretical (Silk 1997 and Elmegreen 1997) arguments and from observational data, $N \sim 1$, and Kennicutt (1998) finds that

$$\Sigma_{SFR} = 0.017 \Omega \Sigma_g^N \tag{14.22}$$

provides just as good a fit to the observational data as the simple Schmidt Law.

Finally Dopita & Ryder (1994) argue that the rate of star formation reflects the cloud-cloud collision timescale in the disk, since this determines the growth timescale of the molecular clouds from which the stars are formed:

$$\Sigma_{SFR} \propto \frac{\Sigma_g}{\tau_{cc}} \tag{14.23}$$

In addition we must also take into account the depth of the local disk gravitational potential, since this determines the maximum pressure that star formation processes can maintain in the disk before the hot phase is vented up galactic chimneys into the halo, and which in turn determines the physical state of the molecular clouds. This leads to a Schmidt-like law involving both the gas surface density, and the total matter surface density, Σ_T:

$$\Sigma_{SFR} = A \frac{\Sigma_T \Sigma_g^{5/3}}{(z_* + z_g)^{1/3} v_{esc}^{2/3}} \tag{14.24}$$

This provides an excellent fit to the observations of normal disk galaxies in which the surface density is estimated using I-band photometry (Ryder & Dopita, 1994).

Clearly, all these formulations of the star formation rates in galaxies cannot be correct. However, they serve to illustrate that there must be a complex feedback between the rate of star formation and the thermal phases of the interstellar medium. One controls the other, which in turn controls the one. We do not yet have a complete understanding of all of the essential physics of this process, but it is clear that it is there, waiting to be elucidated.

14.3.4 Mass Exchange

Mass exchange between the various phases of the interstellar medium and between the interstellar medium and the stars drives the physical and chemical evolution of galaxies. The schematics of this mass exchange is shown in Figure (14.03). The sense of the process is that material from the interstellar medium is constantly being incorporated into low mass stars with $M < 0.8 - 0.9 M_\odot$, which live for longer than a Hubble time, or into stellar remnants; white dwarf stars ($M < 1.4 M_\odot$), neutron stars ($M \sim 1.4 M_\odot$) and black holes ($M > 1.4 M_\odot$). Stellar nucleosynthetic process convert hydrogen and helium into heavier elements, and stellar mass-loss and supernova explosions return some of these heavier elements to the interstellar medium. The initial stellar mass determines both the chemical makeup of this material and the timescale over which it is returned to the diffuse phase.

The massive stars burn their nuclear fuel in a profligate manner, and therefore live only a very short time. Initially they burn hydrogen in their core. A useful approximation to the stellar lifetime in the hydrogen-burning phase is:

$$\left[\frac{\tau}{Myr}\right] \sim 4.5 \left[\frac{M}{40 M_\odot}\right]^{-0.97} ; M \leq 40 M_\odot$$

$$\left[\frac{\tau}{Myr}\right] \sim 4.5 \left[\frac{M}{40 M_\odot}\right]^{-0.43} ; M > 40 M_\odot \qquad (14.25)$$

After these stars have exhausted their hydrogen, they burn helium for a little less than 10^6 yr, then they start on the heavier elements, through $\alpha-$process burning reactions until the core has become an inert mass composed of the iron-peak elements. This contracts, the temperature increases, until the iron is photodissociated, all pressure support is lost and the core collapses to a neutron star or black hole, producing a Type II supernova explosion.

The matter ejected by such explosions is rich in He and the $\alpha-$process elements including O and Si, and the $r-$process elements formed by neutron capture under explosive conditions, but contains only small amounts of Fe and the iron-peak elements. To all intents and purposes this material is returned promptly to the interstellar medium, since the lifetime of the massive stars is a negligible fraction of a galactic evolutionary timescale.

Low mass stars cannot proceed beyond helium burning, and end their lives as mass-losing red giants which then pass through the planetary nebula phase to finally produce a white dwarf. The lifetime of stars is much greater, and is given by equation (10.20):

$$\left[\frac{\tau}{Gyr}\right] \sim 11 \left[\frac{M}{M_\odot}\right]^{-3.1} + 0.46 \left[\frac{M}{M_\odot}\right]^{-4.6} \qquad (14.26)$$

As we discussed in Section (10.4.2), these stars produce an important fraction of the He, ^{12}C, ^{14}N and ^{22}Ne which chemically enriches the interstellar

medium. They also produce the heavy $s-$process elements formed by neutron capture under non-explosive conditions.

If stars of this mass range are found in close binary systems, then mass-exchange between the binary partners in the late phases of stellar evolution can push a degenerate core of one of the stars over the Chandrasekar stability limit of about $M \sim 1.4 M_\odot$, resulting in a thermonuclear supernova, a supernova of Type I, which burns much of the star up to nuclear statistical equilibrium conditions at a temperature in excess of 10^8 K. Such stars produce most of the Fe and Ni found in interstellar gas. However, they do this after a considerable time delay; of order 10^9 yr, which is an important fraction of the gas depletion timescale due to star formation in galaxies. Much more detail about these chemical evolution aspects of the interstellar medium can be found in Pagel (1997).

14.3.5 Dust Evolution in a Multi-Phase Medium

Advanced Topic

We discussed in Section (12.5) how the life-cycle of dust grains in the interstellar medium must, in large measure, represent a balance between grain destruction and shattering in the fast shocks found in the HIM, WIM and WNM phases, and the processes which build up grains in the CNM and in the dense molecular clouds. Dust evolution therefore depends upon the mass transport between these various phases.

Consider, following Dwek & Scalo (1980) and McKee (1989b), the evolutionary equations which describe these processes. Let M is the mass of an element, produced by star-formation and evolution process which can be potentially locked up into a grain material in a given phase of the interstellar medium. Suppose a fraction δ of this is actually locked up into dust in the interstellar medium. The evolution of M will depend on the rate of loss from the ISM due to star formation, occurring over a characteristic timescale τ_{sfr} and the rate of injection to the ISM from evolved stars, occuring over a timescale τ_{in} . Ignoring the rate of mass transport between the various phases:

$$\frac{dM}{dt} = -\frac{M}{\tau_{sfr}} + \frac{M}{\tau_{in}}$$

(14.27)

The fraction of this mass locked into dust will depend upon the fraction α destroyed or manufactured ($\alpha < 1$) in star-forming regions and the rate of star formation. It will also depend on the injection rate, and the rate at which dust is destroyed by supernova shocks, occurring over a characteristic timescale τ_{SNR}:

$$\frac{d\,(M\delta)}{dt} = -\frac{\alpha M\delta}{\tau_{sfr}} + \frac{M\delta_{in}}{\tau_{in}} - \frac{M\delta}{\tau_{SNR}}$$

(14.28)

From these two equations the dust fraction evolves as:
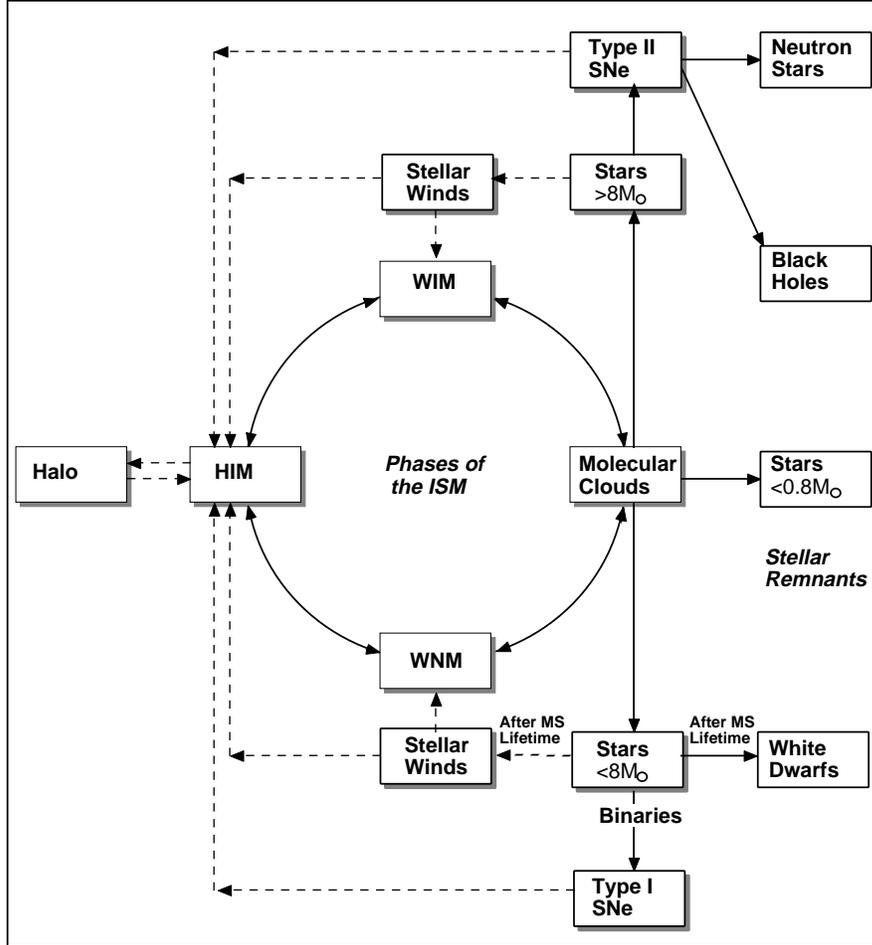
Fig. 14.3. The star-gas cycle in galactic diffuse media.

$$\frac{d\delta}{dt} = -\frac{(\alpha - 1)\,\delta}{\tau_{sfr}} + \frac{(\delta_{in} - \delta)}{\tau_{in}} - \frac{\delta}{\tau_{SNR}} \qquad (14.29)$$

and will eventually reach an equilibrium value:

$$\delta_{eq} = \delta_{in} \left[1 + \frac{\tau_{in}}{\tau_{SNR}} + (\alpha - 1)\,\frac{\tau_{in}}{\tau_{sfr}}\right]^{-1} \qquad (14.30)$$

In this case the equilibrium logarithmic depletion factor for the element concerned is $\log\left[1 - \delta_{eq}\right]$. In the hot ISM, $\tau_{SNR} \ll \tau_{in}$, so the depletion factors will be small, while material that has been incorporated into dense clouds will display a depletion factor which increases with time towards the equilibrium

value at a rate which depends on the cloud density. This is precisely what is seen in EUV absorption line measurements using the *Copernicus* satellite by Jenkins (1987; 1989).

If we allow for mass transport for a simplified two-phase medium consisting of a dense cold and a tenuous hot medium, we can assume that the timescale to destroy the dust though supernova shocks in the dense phase is much longer than the time needed to destroy the dust by shocks in the low-density phase. If we also set $\alpha = 1$ then the equilibrium dust fraction in the cold and in the hot phases, δ_C and δ_H, respectively, are given by:

$$\frac{\delta_C - \delta_H}{\delta_C} = \frac{1/\tau_{SNR}}{(1/\tau_{C \to H} + 1/\tau_{C \to H} + 1/\tau_{in} + 1/\tau_{SNR})} \sim \frac{\tau_{C \to H}}{\tau_{SNR}} \qquad (14.31)$$

In other words, the difference in the dust fractions between the two phases is simply the ratio of the time taken for the dust to escape into the hot phase to the time taken to destroy it in the hot phase.

Notes on Chapter 14

- Much of the physics of the multiphase interstellar medium described in this chapter is drawn from the reviews of McKee, C.F. 1995, in *The Physics of the Interstellar & Intergalactic Medium*, eds. A. Ferrara, C.F. McKee, C. Heiles & P.R. Shapiro, ASP Conf. Ser. 80, 292 and of Begelman, M.C. 1990, in *The Interstellar Medium in Galaxies*, eds H.R. Thronson & J.M. Shull, Kluwer:Dordrecht, p287 (ISBN 0-7923-0759-3). See also, the seminal papers by McKee, C.F. and Ostriker, J.P. 1977, ApJ, 218, 148 and by Begelman, M.C. & McKee, C.F. 1990, ApJ, 358, 375.

Exercises

1

2

3

1. Seaton , M. J. 1958, Rev. Mod. Phys, 30, 979.
2. Landini, M. and Fosse, B. C. M. 1990, A&A Suppl., 82, 229.

4

1. Allen, D.A. 1980, MNRAS, 190, 75.
2. Baan, W.A. 1985, Nature, 315, 26.
3. Baan, W.A. 1989, ApJ, 338, 804.
4. Baan, W.A. & Haschick, A.D. 1984, ApJ, 318, 139.
5. Braatz, J.A., Wilson, A.S. & Henkel, C. 1994, ApJ, 437, L99.
6. Braatz, J.A., Wilson, A.S. & Henkel, C. 1996, ApS, 106, 51.
7. Cooke, B. & Elitzur, M. 1985, ApJ, 295, 175.
8. Curiel, S., Raymond, J.C., Wolfire, M., Hartigan, P. Morse, J., Schwartz, R.D., & Nisenson, P. 1995, ApJ, 453, 322.
9. Deguchi, S. 1977, PASJ, 29, 669.
10. Drake, S.A. & Ulrich, R.K. 1980, ApJS, 42, 351.
11. Dring, A.R., Linsky, J., Muthy, J. Henry, R.C., Moos, W., Vidal-Majar, A., Auduoze, J., & Landsman, W. 1997, ApJ, 488, 760.
12. Elitzur, M. 1976, ApJ, 203, 124.
13. Elitzur, M. 1978, A&A, 62, 305.
14. Elitzur, M. Hollenbach, D.J. & McKee, C.F. 1989, ApJ, 346, 983.
15. Greenhill, L.J., Gwinn, C.R., Antonucci, R. & Barvainis, R. 1996, ApJ, 472, L21.
16. Greenhill, L.J., Jiang, R.D., Moran, J.M., Reid, M.J., Lo, K.Y., & Claussen, M.J., 1995, ApJ, 440, 619.
17. Harvey, P.M., Bachis, K.P., Wilson, W.J., & Ball, J.A. 1974, ApJS, 27, 331.
18. Jordan *et al.* 1977, Nature, 270, 326.
19. Koekemoer, A.M., Henkel, C., Greenhill, L.J., Dey, A., van Breugel, W., Codella, C. & Anonucci, R. 1995, Nature, 378, 697.
20. Linsky, J.L, Diplas, A., Wood, B.E., Brown, A., Ayres, T.R., & Savage, B.D. 1995, ApJ, 451, 335.
21. Maloney, P.R. 1997, in *Accretion Phenomena and Related Outflows*, eds. D.T. Wickramasinghe, L. Ferrario & G.V. Bicknell, ASP Conf. Ser, v121 (Ast Soc Pacific : San Francisco), p113.
22. Melnick, G.J., Menten, K.M., Phillips, T.G. & Hunter, T. 1993, ApJ, 416, L37.
23. Miyoshi, M., Moran, J., Herrnstein, J., Greenhill, L., Nakai, N., Diamond, P. & Inoue, M. 1995, Nature, 373, 127.
24. Neufeld, D.A. & Melnick, G.J. 1991, ApJ, 368, 215.
25. Palma, A., Green, S., DeFres, D.J., & McLean, A.D. 1988, ApJS, 68, 287.
26. Savage, B.D. & Sembach, K.R. 1996, Ann.Rev.A&A, 34, 279.
27. Schmid, H.M. 1989, A&A, 211, 31.
28. Schwartz, R.D. 1983, Ann. Rev. Ast. Ap., 21, 209.
29. Spitzer, L. & Fitzpatrick, E.L. 1993, ApJ, 409, 299.

5

1. Aldrovandi, S.M.V. & Gruenwald, R.D. 1985, A&A, 147, 331.
2. Arnaud, M. & Rothenflug, R. 1985, A&AS, 60, 425.
3. Band, I.M., Trzhaskovskaya, M.B., Verner, D.A., & Yakovlev, D.G. 1990, A&A, 237, 267.
4. Berrington, K. A., Burke, P. G., Butler, K., Seaton, M. J., Storey, P. J., Taylor, K. T., & Yan, Yu. 1987, J. Phys. B, 20, 6379
5. Hummer, D.G., Berrington, K.A., Eissner, W., Pradhan, A.K., Saraph, H.E. & Tully, J.A. 1993, A&A, 279, 298.
6. McCarroll, R. & Valiron, R. 1976, A&A, 53, 83.
7. Nahar, S.N., & Pradhan, A.K. 1994, Phys Rev. A, 49, 1816.
8. Nahar, S.N., & Pradhan, A.K. 1997, ApJS, 111, 339.
9. Petrini, D & Da Silva, F.P. 1997, A&A, 317, 262.
10. Petrini, D & De Araújo, F.X. 1994, A&A, 282, 315.
11. Petrini, D & Farcas, G.A.. 1994, A&A, 292, 337.
12. Shull, J.M & Van Steenberg, M.E. 1985, ApJ, 298, 268.
13. Sutherland, R. S. & Dopita, M.A. 1993, ApJS, 88, 253.
14. Younger, S.M. 1981, J. Quant. Spect & Rad. Trans, 26, 329.
15. Younger, S.M. 1982, J. Quant. Spect & Rad. Trans, 27, 541.
16. Younger, S.M. 1983, J. Quant. Spect & Rad. Trans, 29, 61.

6

1. Breit, C. & Teller, E. 1940, ApJ, 91, 215.
2. Brocklehurst, M. 1971, MNRAS, 153, 471.
3. Brocklehurst, M. 1972, MNRAS, 157, 211.
4. Brocklehurst, M. & Seaton, M.J. 1972, MNRAS, 157, 211.
5. Brown, R.L. 1987, in *Spectroscopy of Astrophysical Plasmas*, eds. A. Dargarno & D. Layzer, CUP:Cambridge, p. 35-58. (ISBN 0 512 23615 8)
6. Dopita, M.A., Binette, L. & Schwartz, R.D., 1982, ApJ, 261, 183.
7. Goldberg, L. 1966, ApJ, 144, 1225.
8. Spitzer, L., Jr. & Greenstein, J.L. 1951, ApJ, 114, 407.

7

1. Begelman, M.C. & Fabian, A.C. 1990, MNRAS, 244, 26p.
2. Begelman & McKee, 1990, ApJ, 358, 375.
3. Bohlin, R.R., Savage, B.D., & Drake, J.F. 1978, ApJ, 224, 132.
4. Cayatte, V., van Gorkon, J.H., Balkowski, C. & Kotanyi, C. 1989, AJ, ****.
5. Chevalier, R.A. & Oegerle, W.R. 1979, ApJ, 277, 398.
6. Ciotti, L, D'Ercole, A., Pellegrini, S., & Renzini, A. 1991, ApJ, 376, 380.
7. Cowie, C.F. & McKee, C.F. 1977, ApJ, 211, 135.
8. Dickey, J.M. & Lockman, F.J. 1990, Ann. Rev. A&A, 28, 215.
9. Diplas, A. & Savage, B.D. 1994, ApJ, 427, 274.
10. Fabbiano, G., Kim, D-W., & Trinchieri, G. 1992, ApJS, 80, 531.
11. Giovanelli, R. & Haynes, M.P. 1985, ApJ, 292, 404.
12. McCray, R. 1987, in *Spectroscopy of Astrophysical Plasmas,* eds. A. Dalgarno & D. Layzer, (CUP:Cambridge), p255.
13. McKee, C.F. & Cowie, L.L. 1977, ApJ, 215, 213.
14. Mathewson, D.S., & Ford, V.L, 1984 in *Structure and evolution of the Magellanic Clouds* (ISBN 85-26576 11-90). (Reidel: Dordrecht), p.125.
15. Meurer, G.R., Bicknell, G.V. & Gingold, R.A. 1985, PASA, 6, 195.
16. Moore, B. & Davis, M. 1994, MNRAS, 270, 209.
17. Nordgren, T.E., Cordes, J.M., & Terzian, Y. 1992, AJ, 104, 1465.
18. Reynolds, R.J. 1980, ApJ, 236, 153.
19. Reynolds, R.J. 1993 in Back to the Galaxy, eds S.G. Holt & F. Verter (AIP:New York), p156.
20. Sarazin, C.L. & McNamara, B.R. 1997, ApJ, 480, 203.
21. Savage, B.D. in *The Physics of the Interstellar Medium and Intergalactic Medium*, ASP Conf. Ser. v80; eds A. Ferrara, C.F. McKee, C.Heiles & P.R. Shapiro, (ASP:SanFrancisco), p233 (ISBN 0-9377-7-99-6).
22. Sembach, K.R. & Savage, B.D. 1992, ApJS, 83, 147.
23. Slavin, J.D., Shull, J.M. & Begelman, M.C. 1993, ApJ, 407, 83.
24. Shapiro, P.R. & Benjamin, R.A. 1993, in *Star Formation, Galaxies and the Interstellar Medium*; eds J.Franco, F. Ferrini and G. Tenorio-Tagle, (CUP: Cambridge), p275 (ISBN 0-521-444128).
25. Shapiro, P.R. & Field, G.B. 1976, ApJ, 205, 762.
26. Shull, J.M. & Slavin, J.D. 1994, ApJ, 427, 784.
27. Spitzer, L. Jr. 1996, ApJ, 458, L29.
28. Sutherland, R.S. & Dopita, M.A. 1993, ApJS, 88, 253.
29. Wang, Q.D, & McCray, R. 1993, ApJ, 409,L37.
30. Weiner, B.J. & Williams, T.B. 1996, AJ, 111, 1156.

8

1. Blair, W.P. & Kirshner, R.P. 1985, ApJ, 298, 582.
2. Böhm, K.-H. & Goodson, A.P. 1997, in *Herbig-Haro Flows & the Birth of Low Mass Stars*, eds. B. Reipurth & C. Bertout, Kluwer:Dordrecht, p47.
3. Castor, J., McCray, R. & Weaver, R. 1975, ApJL, 200, L107.
4. Chernoff, D.F. 1987, ApJ, 312, 143.
5. Chevalier, R.A. & Raymond, J.C. 1978, ApJ, 225, L27.
6. Cowie, L.L. & McKee, C.F. 1977, ApJ, 211, 135.
7. Cox, D.P. 1972, ApJ, 178, 143.
8. D'Odorico, S., Dopita, M.A., & Benvenuti, P. 1980, ApJ, 236, 628.
9. Dopita, M.A. 1976, ApJ, 209, 395.
10. Dopita, M.A. 1977, ApJS, 33, 437.
11. Dopita, M.A., Binette, L. 1983, in IAU Symp#101, *Supernova Remnants and their X-Ray Emissions*, eds. J.Danziger & P. Gorenstein, Reidel:Dordrecht, p221.
12. Dopita, M.A., Binette, L., D'Odorico, S. & Benvenuti, P. 1984, ApJ, 276, 653.
13. Dopita, M.A., Binette, L. & Schwartz, R.D., 1982, ApJ, 261, 183.
14. Draine, B.T. 1980, ApJ, 241, 1021.
15. Draine, B.T. , Roberge, W.G. & Dalgarno, A. 1983, ApJ, 264, 485.
16. Fesen, R.A., Blair, W.P., & Kirshner, R.P. 1985, ApJ, 262, 171.
17. Flower, D.R., Pineau des Forêts, G., Field, D. & May, P.W. 1996, MNRAS, 280, 447.
18. Heiles, C., Goodman, A.A., McKee, C.F. & Zweibel, E.G. 1993 in *Protostars and Planets III*, ed E.H. Levy, J.I. Lunine, & M.S. Matthews, U.Ariz. Press: Tucson, p279.
19. Hester, J.J., Raymond, J.C., & Danielson, G.A. 1986, ApJ, 303, L17.
20. Laming, J.M., Raymond, J.C., McLaughlin, B.M., & Blair, W.P. 1996, ApJ, 472, 267.
21. McKee, C.F. & Cowie, L.L. 1977, ApJ, 215, 213.
22. McKee, C.F. & Ostriker, J.P. 1977, ApJ, 218, 148.
23. Mathewson, D.L. *et al.* 1983, ApJS, 51, 345.
24. Nadeau, D. & Geballe, T.R. 1979, ApJ, 230, L169.
25. Payne, D. G. & Cohn, H. 1985, ApJ, 291, 655.
26. Raymond, J.C. 1979, ApJS, 39, 1.
27. Schweitzer, F. & Lasker, B. M. 1978, ApJ, 226, 167.
28. Sellmaier, F. *et al.* 1993, A&A, 273, 533.
29. Smith, M.D. 1993, ApJ, 390, 447.
30. Smith, M.D. & Brand, P.W.J.L. 1990, MNRAS, 242, 495.
31. Smith, M.D. & Mac Low, M-M. 1997, A&A, 326, 801.
32. Shull, J.M. & McKee, C.F. 1979, ApJ, 227, 131.
33. Shu, F.H. *et al.* 1994, ApJ., 429, 781.

338   8

34. Tuohy, I.R., Dopita, M.A., Mathewson,D.S., Long, K.S., & Helfand, D.J. 1982, ApJ, 261, 473.
35. Weaver, R., McCray, R., Castor, J., Shapiro, P. & Moore, R. 1977, ApJ, 218, 377.
36. Winkler, P.F., Bromley, B.C. & Canizares, C.R. 1983, in IAU Symp#101, *Supernova Remnants and their X-Ray Emission*, eds. J.Danziger & P. Gorenstein, Reidel:Dordrecht, p245.

9

1. Baker, J.G. & Menzel, D.H. 1938, ApJ, 88, 52.
2. Brocklehurst, M. 1971, MNRAS, 153, 471.
3. Brocklehurst, M. 1972, MNRAS, 157, 211.
4. Brocklehurst, M., & Seaton, M.J. 1972, MNRAS, 157, 179.
5. Dyson, J.E. 1968, Ap.& Sp. Sci., 1, 388.
6. Ferland, G.J. 1993, Internal Report, Univ. of Kentucky, dept. of Physics & Astronomy.
7. Ferland, G.J., & Netzer, H. 1983, ApJ, 264, 105.
8. Goldberg, L. 1966, ApJ, 144, 1225.
9. Kahn, F.D. 1954, BAN, #465.
10. Koski, A.T. 1978, ApJ, 223, 56.
11. MacAlpine, G.M. 1972, ApJ, 175, 11.
12. Menzel, D.H. & Baker, J.B. 1937, ApJ, 86, 70.
13. Murayama, T. & Taniguchi, Y. 1998, ApJ, 497, L9.
14. Seaton, M.J. 1964, MNRAS, 127, 177.
15. Seaton, M.J. 1978, MNRAS, 185, 5P.
16. Stasinska, G. 1984, A&A, 135, 341.
17. Strömgren, B. 1939, ApJ, 89, 526.
18. Sutherland, R.S. & Dopita, M.A. 1993, ApJS, 88, 253.
19. Williams, R.E. 1967, ApJ, 147, 556.
20. Wilson, O.C., Münch, G., Flather, E. & Coffeen, M. 1959, ApJS, 4, 199.

10

1. Afflerbach, A. *et al.* 1996, ApJS, 106, 423.
2. Afflerbach, A., Churchwell, E. & Werner, M.W. 1997, ApJ, 478, 190.
3. Anders, E. & Grevesse, N. 1989, Geochim. Cosmochim. Acta, 53, 197.
4. Baldwin, J.A., Phillips, M.M., and Terlevich, R. 1981, PASP, 93, 5.
5. Binette, L., Dopita, M.A., D'Odorico, S., & Benvenuti, P. 1982, A&A, 115, 315.
6. Brown, R.L. 1987, in *Spectroscopy of Astrophysical Plasmas* eds. A.Dalgarno & D. Layzer, CUP: Cambridge, p35.
7. D'Odorico, S. 1978, Mem.SA It, 49, 485.
8. Dopita, M.A., 1972, A&A,17, 165.
9. Dopita, M.A. 1977, ApJS, 33, 437
10. Dopita, M.A. 1978, ApJS, 37, 117.
11. Dopita, M.A. & Evans, I.N. 1986, ApJ, 307, 431.
12. Dopita, M.A., Gibbons, A.H. and Meaburn, J., 1973, A&A, 22, 33.
13. Dopita, M.A. *et al.* 1997, ApJ, 474, 188.
14. Edmunds, M.G. & Pagel, B.E.J. 1984, MNRAS, 211, 507.
15. Friedli, D. & Benz, W. 1995, A&A, 301, 649.
16. French, H.B. & Grandi, S.A. 1981, ApJ, 244, 493.
17. Gabler, R., Gabler, A., Kudritzki, R.P., Puls, J., Pauldrach, A.W.A 1989, A&A, 226, 162.
18. Garnett, D. & Shields, G.A. 1987, ApJ, 317, 82.
19. Gabler, R., Kudritzki, R.P., & Mendez 1991, A&A, 245, 587.
20. Götz, M. & Köppen, J., 1992, A&A, 262, 455.
21. Gruenwald, R. & Viegas, S.M. 1995, A&A, 303, 535.
22. Gruenwald, R., Viegas, S.M., & Broguière, D. 1997, ApJ, 480, 283.
23. Gruenwald, R. & Viegas, S.M. 1998, ApJ (in press).
24. Iben, I. Jr., and Renzini, A. 1983, A&A, 21, 271.
25. Jeffries, R.D. 1997, MNRAS, 288, 585.
26. Kaler, J.B. 1976, ApJ, 210, 843.
27. Kaler, J.B. 1989 in Iau Symp#131: *Planetary Nebulae*, ed. S. Torres-Peimbert, Kluwer:Dordrecht, p229.
28. Kennicutt, R.C. Jr. & Garnett, D.R. 1996, ApJ, 456, 504.
29. Kwitter, K.B. & Henry, R.B.C. 1998, ApJ, 493, 247.
30. Herrero, A., Machado, A., & Méndez, R.H., 1990, Ap&SS, 169, 183.
31. McCall, M.L., Rybski, P.M. & Shields, G.A. 1985, ApJS, 57, 1.
32. Maciel, W.J. & Köppen, J. 1994, A&A, 282, 43.
33. Marigo, P., Bressan, A. and Chiosi, C. 1996, A&A, 313, 545.
34. Mathis, J.S., Torres-Peimbert, S. & Peimbert, M. 1998, ApJ, 495, 328.
35. Matteucci, F. & Francois, P. 1989, MNRAS, 239, 885.
36. Méndez, R.H., Kudritzki, R.P., Herrero, A., Husfeld, D., & Groth, H.G. 1988, A&A, 190, 113.
37. Meyer, D. M., Jura, M. & Cardelli, J.A. 1998, ApJ, 493, 222.

38. Osterbrock, D.E., Tran, H.D., and Veilleux, S. 1992, ApJ, 389, 196.
39. Pagel, B.E.J., Edmunds, M.G., Blackwell, D.E., Chun, M. S. & Smith, G. 1979, MNRAS, 189, 95.
40. Peimbert, M. & Costero, R. 1969, Bol. Obs. Tonantzintla y Tacubaya, 5, 3.
41. Peimbert, M., Storey, P.J., & Torres-Peimbert, S. 1993, ApJ, 414, 626.
42. Peimbert, M. & Torres-Peimbert, S. 1977, MNRAS, 179, 217.
43. Prantzos, N. & Aubert , O. 1995, A&A, 302, 69. and
44. Preite-Martinez, A. & Pottasch, A.S. 1983, A&A, 126, 31.
45. Ratag, M.A., Pottasch, S.R., Dennefeld, M. & Menzies, J. 1997, A&AS, 126, 297.
46. Renzini A. & Voli, M. 1981, A&A, 94, 175.
47. Roy, J.-R. & Walsh, J.R. 1997, MNRAS, 288, 715.
48. Sabbadin, F., Minello, S., and Bianchini, A. 1977, A&A, 60, 147
49. Sanduleak, N. 1969, AJ, 74, 47.
50. Schwarz, G.J., Starrfield, S., Shore, S.N., & Hauschildt, P.H. 1997, MNRAS, 290, 75.
51. Searle, L. 1971, ApJ, 168, 327.
52. Shaver, P.A., McGee, R.X., Newton, L.M., Danks, A.C. & Pottasch, S.R. 1983, MNRAS, 204, 53.
53. Simpson, J.P. *et al.* 1995, ApJ, 444, 721.
54. Smartt, S.J. & Rolleston, W.R.J. 1997, ApJ, 481, L47.
55. Smits, D.P. 1991, MNRAS, 248, 193.
56. Stasinska, G. 1978, A&A, 66, 257.
57. Stasinska, G. 1980, A&A, 84, 320.
58. Stoy, R.H. 1933, MNRAS, 93, 588.
59. Veilleux, S., and Osterbrock, D.E. 1987, ApJ S., 63, 295.
60. Vanlandingham, K.M., Starrfield, S., Wagner, R.M., Shore, S.N., & Sonneborn, G. 1996, MNRAS, 212, 753.
61. Vanlandingham, K.M., Starrfield, S., & Shore, S.N. 1997, MNRAS, 290, 87.
62. Vanlandingham, K.M., Starrfield, S., Shore, S.N., & Sonneborn, G. 1998, MNRAS, (in press).
63. Vassiliadis, E., & Wood, P.R. 1993, ApJ, 413, 641.
64. Walsh, J.R. & Roy, J.-R. 1997, MNRAS, 288, 726.
65. Williams, R.E. 1982, ApJ, 261, 170.
66. Williams, R.E., Woolf, N.J., Hege, E.K., Moore, R.L. & Kopriva, D.A. 1978, ApJ, 224, 171.
67. Zanstra, H. 1931, Publ. Dom. Astrophys. Obs., 4, 209.

11

1. Allen, M.G., Dopita, M.A. & Tsvetanov, Z.I. 1998, ApJ, 493, 571.
2. Blair, W.P., Kirshner, R.P. & Winkler, P.F. 1983, ApJ, 272, 84.
3. Blair, W.P., Raymond, J.C., Danziger, J. & Matteucci, F. 1989, ApJ, 338, 812.
4. Binette, L., Dopita, M.A. & Tuohy, I.R. 1985, ApJ, 297, 476.
5. Chevalier, R.A. & Kirshner, R.P. 1979, ApJ, 233, 154.
6. Cox, D.P. & Raymond, J.C. 1985, ApJ, 298, 651.
7. De Kool, M., Li, H. & McCray, R. 1998, ApJ, 503, 857.
8. Daltabuit, E. & Cox, D. 1972, ApJ, 173, L13.
9. Dopita, M.A. 1987, Aus. J. Phys., 40, 789.
10. Dopita, M.A. *et al.* 1997, ApJ, 490 202.
11. Dopita, M.A., Binette, L. & Schwartz, R.D. 1982, ApJ, 161, 183.
12. Dopita, M.A., Binette, L. and Tuohy, I.R. 1984, ApJ, 282, 142.
13. Dopita, M.A., Tuohy, I.R 1984, ApJ, 282, 135.
14. Dopita, M.A., Tuohy, I.R. & Mathewson, D.S. 1981, ApJ, 248, L105.
15. Dopita, M.A. & Sutherland, R.S. 1995, ApJ, 455, 468.
16. Dopita, M.A. & Sutherland, R.S. 1996, ApJS, 102, 161.
17. Ferland, G.J. & Netzer, H. 1983, ApJ, 264, 105.
18. Goss, W.M. *et al.* 1979, MNRAS, 188, 357.
19. Heckman, T.M. 1980, A&A, 87, 142.
20. Ho, L., Fillippenko, A.V. & Sargent, W.L.W. 1993, ApJ, 417, 63.
21. Ho, L., Fillippenko, A.V. & Sargent, W.L.W. 1995, ApJS, 98, 477.
22. Huang, Z.P. *et al.* 1994, ApJ, 424, 114.
23. Itoh, H. 1981a, PASJ, 33, 121.
24. Itoh, H. 1981b, PASJ, 33, 521
25. Kamper, K.W. and van den Bergh, S. 1976, ApJS, 32, 351.
26. Kirshner, R.P. *et al.* 1989, ApJ, 342, 260.
27. Koski, A.T. 1978, ApJ, 223, 56.
28. Kriss *et al.* 1992, ApJ, 394, L37.
29. Lasker, B.M. 1978, ApJ, 233, 109.
30. Long, K.S., Helfand, D.J. & Grabelsky, D.A. 1981, ApJ, 248, 925.
31. Morse *et al.* 1996, AJ, 112, 509.
32. Muxlow, T.W.B. *et al.* 1995, MNRAS, 266, 455.
33. Osterbrock, D.E. Tran, H.D. & Veilleux, S. 1992, ApJ, 389, 196.
34. Phillips, M.M. *et al.* 1986, AJ, 91, 1062.
35. Raymond, J.C. 1976, Thesis, U. Wisconsin.
36. Shull, J.M. & McKee, C.F. 1979, ApJ, 227, 131.
37. Shull, P. Jr. 1983, ApJ, 275, 611.
38. Stasinska, G. 1984, A&A, 135, 341.
39. Sutherland, R.S., Bicknell, G.V. & Dopita, M.A. 1993, ApJ, 414, 510.
40. Sutherland, R.S. and Dopita, M.A. 1995a, ApJ, 439, 365.
41. Sutherland, R.S. and Dopita, M.A. 1995b, ApJ, 439, 381.

42. Veilleux, S. & Osterbrock, D.E. 1987, ApJS, 63, 295.
43. Viegas-Aldrovandi, S.M. & Contini, M. 1989, ApJ, 339, 689.
44. Winkler, P.F. & Kirshner, R.P. 1985, ApJ, 299, 981.

12

1. Aitken, D. K. & Roche, P. F. 1984, MNRAS, 208, 751.
2. Allamandola, L., Tielens, A.G.G.M., & Barker, J.R. 1989, ApJS, 71,733.
3. Bianchi, L., Clayton, G.C., Bohlin, R.C., Hutchings, J.B. & Massey, P. 1996, ApJ, 471, 203.
4. Bohlin, R.C., Savage, B.D. & Drake, J.F. 1978, ApJ, 224, 132.
5. Bregman, J.D., Campins, H., Witteborn, F.C., Wooden, D.H., Rank, D.M., Allamandola, L. J., Cohen, M. & Tielens, A.G.G.M. 1987, A&A, 187, 616.
6. Cherneff, I., Barker, J.R., & Tielens, A.G.G.M. 1992, ApJ, 401, 269.
7. Cohen, M., Allamandola, L.J., Tielens, A.G.G.M., Bregman, J.D., Simpson, J.P., Witteborn, F.C., Wooden, D.H., & Rank, D. 1986, ApJ, 302, 737.
8. Cohen, M., Tielens, A.G.G.M., Bregman, J.D., Witteborn, F.C., Rank, D., Allamandola, L. J., Wooden, D. H., & de Muizon, M. 1989, ApJ, 341, 246.
9. Desert, F.X., Boulanger, F. & Shore, S.N. 1986, A&A, 160, 295.
10. Draine, B.T. 1978, ApJS, 36,595.
11. Draine, B.T. 1988, ApJ, 333, 848.
12. Draine, B.T. & Anderson, N. 1985, ApJ, 292, 494.
13. Draine, B.T. & Salpeter, E.E. 1979, ApJ, 231, 77.
14. Draine, B.T. & Sutin, B. 1987, ApJ, 320, 803.
15. Duley, W.W. 1973, Ap. Sp. Sci., 23, 43.
16. Dwek, E., & Arendt, 1992, Ann. Rev. A&A, 30, 11.
17. Dwek, E., Foster, S.M., & Vancura, O. 1996, ApJ, 457, 244.
18. Dwek, E., *et al.* 1997, ApJ, 474, 565.
19. Fitzpatrick, E.L. 1985, ApJ, 299, 219.
20. Fitzpatrick, E.L. 1986, AJ, 92, 1068.
21. Fitzpatrick, E.L. & Massa, D. 1986, ApJ, 307, 286.
22. Fitzpatrick, E.L. & Massa, D. 1988, ApJ, 328, 734.
23. Fitzpatrick, E.L. & Massa, D. 1990, ApJS, 72, 163.
24. Frenklach, M. & Feigelson, E.D. 1989, ApJ, 341, 372.
25. Guhathakurta, P., & Draine, B.T. 1989, ApJ, 345, 230.
26. Habing, H.J., Tignon, J. & Tielens, A.G.G.M. 1994, A&Ap, 286, 523.
27. d'Hendcourt, Léger, A., Boissel, P. & Désert, F.X. 1989,in *Interstellar Dust*, eds. L.J. Allamandola & A.G.G.M. Tielens, Kluwer:Dordrecht, p207.
28. Hibbins *et al.* 1995, , in *The Diffuse Interstellar Bands*, eds. A.G.G.M. Tielens & T.P. Snow Klwer: Dordrecht, p25.
29. Hollenbach, D. & Salpeter, E.E. 1970, J. Chem. Phys., 53, 79.
30. van der Hulst, H. C. 1957, *Light Scattering by Small Particles,* Wiley: New York.
31. Jenniskens, P & Désert, F.-X. 1995, in *The Diffuse Interstellar Bands*, eds. A.G.G.M. Tielens & T.P. Snow Klwer: Dordrecht, p39.
32. Jones, A.P., Tielens, A.G.G.M., Hollenbach, D.J. & McKee, C.F. 1994, ApJ, 433, 797.

33. Jones, A.P., Tielens, A.G.G.M., & Hollenbach, D.J. 1996, ApJ, 469, 740.

34. Kozasa, T. & Sogawa, H. 1997, Ast.&Sp.Sci., 251, 165.

35. Krełowski, J. & Walker, G.A.H. 1987, ApJ, 312, 680.

36. Lacy *et al.* 1984, ApJ, 276, 533.

37. Léger, A. & Puget, J.-L. 1984, A&A, 137, L5.

38. Léger, A., d'Hendecourt, L.B., Verstraete, L., & Ehrenfreund, P. 1991, in *Chemistry in Space*, eds. J.M. Greenberg & V. Pirronello, Kluwer: Dordrecht, p211.

39. Leitch-Devlin, M. A., & Williams, D. A 1984, MNRAS, 213, 295.

40. Loup, C., Foreveille, T., Omont, A., & Paul, J.F. 1993, A&ApS, 99, 291.

41. Maciel, W.J., & Pottasch, S.R. 1982, A&A, 106, 1.

42. Mathis, J.S., Mezger, P.G., & Panagia, N. 1983, A&A, 128, 212.

43. Mathis, J.S., Rumpl, W., & Nordsieck, K.H. 1977, ApJ, 217, 425.

44. Miles, J.R., Sarre, P.J. & Scarrott, S.M. 1995, in *The Diffuse Interstellar Bands*, eds. A.G.G.M. Tielens & T.P. Snow Klwer: Dordrecht, p143.

45. Nuth, J.A. & Moore, M.H. 1988, ApJ, 329, L113.

46. Oliveira, S. & Maciel, W.J. 1986, Ap. Sp. Sci., 126, 211.

47. Omont, A., 1991, in *Chemistry in Space*, eds. J.M. Greenberg & V. Pirronello, Kluwer: Dordrecht, p171.

48. Petrosian, V., Silk, J., & Field, G.B. 1972, ApJ, 177, L69.

49. Pineau des Forêts, G. & Flower, D. 1997, in IAU Symp. #178, *Molecules in Astrophysics: Probes & Processes*, ed. E.F. van Dishoeck, Kluwer:Dordrecht, p113.

50. Prévot, L., Lequeux, J., Maurice, E., Prévot, L., & Rocca-Volmerange, B. 1984, A&A, 132, 389.

51. Purcell, E.M. 1976, ApJ, 206, 685.

52. Rodrigues, C.V., Magalhaes, A.M., Coyne, G.V.S. & Pirola, V. 1997, ApJ, 485, 618.

53. Sandford, S.A. 1989, in *Interstellar Dust*, eds. L.J. Allamandola & A.G.G.M. Tielens, Kluwer:Dordrecht, p403.

54. Sarazin, C.L. 1975, ApJ, 211, 772.

55. Sauvage, M. 1998, ****

56. Schwering, P.B.W., & Israel, F.P. 1991, A&A, 246, 231.

57. Sellgren, K., Allamandola, L.J, Bregman, J., Werner, M.W., & Wooden, D.H. 1985, ApJ, 299, 416.

58. Spitzer, L. Jr. 1976, Comments Astrophys, 6, 177.

59. Tielens, A.G.G.M 1989, in *Interstellar Dust*, eds. L.J. Allamandola & A.G.G.M. Tielens, Kluwer:Dordrecht, p239.

60. Tielens, A.G.G.M. & Whittet, D.C.B. 1997, in IAU Symp. #178, *Molecules in Astrophysics: Probes & Processes*, ed. E.F. van Dishoeck, Kluwer:Dordrecht, p45.

61. Tielens, A.G.G.M., Allamandola, L.J, Bregman, J., Goebel, J., d'Hendecourt, L.B., & Witteborn, F.C. 1984, ApJ, 287, 697.

62. Tielens, A.G.G.M. & Allamandola, L.J. 1987, in *Interstellar Processes*, eds. J. Hollenbach & H.A. Thronson Jr. Kluwer: Dordrecht, p239.

63. Waters, L.B.F.M. *et al.* 1996, A&A, 315, L361.

64. Whittet, D.C.B., Martin, P.G., Hough, J.H., Rouse, M.F., Bailey, J.A. & Axon, D.J. 1992, ApJ, 386, 562.

65. Whittet, D.C.B. & van Breda, I.G. 1978, A&A, 66, 57.

13

1. Boland, W. & de Jong, T. 1982, ApJ, 261, 110.
2. Dalgarno, A. & Lepp, S. 1984, ApJ, 287, L47.
3. Dalgarno, A. & McCray, R. A. 1972, Ann. Rev. A&A, 10, 375.
4. Elizur, M., Hollenbach, D.J. & McKee, C.F. 989, ApJ, 346, 983.
5. Geballe, T.R. & Oka, T. 1996, Nature, 384, 334.
6. Gensheimer, P.D., Mauersberger, R. & Wilson, T.L. A&A, 314, 281.
7. Keene, J., Scilke, P., Kooi, J., Lis, D.C., Mehringer, D.M. & Phillips, T.G. 1998, ApJ, 494, L107.
8. d'Hendecourt, L.B., Allamandola, L.J., Baas, F. & Greenberg, J.M. 1982, A&A, 109, L12.
9. Herbst, E. & Klemperer, W. 1973, ApJ, 185, 505.
10. Hollenbach, D.J. & Salpeter, E.E. 1971, ApJ, 163, 155.
11. Lee, H-H., Herbst, E., Pineau de Fôrets, G., Roueff, E. & LeBourlot, J. 1996, A&A, 311, 690.
12. Leger, A., Jura, M. & Omont, A. 1985, A&A, 144, 147.
13. Lepp, S. & Dalgarno, A. 1996, A&A, 306, L21.
14. Maloney, P.R. 1997, in *Accretion Phenomena and Related Outflows*, eds. D.T. Wickramasinghe, L. Ferrario & G.V. Bicknell, ASP Conf. Ser, v121 (Ast Soc Pacific : San Francisco), p113.
15. Maloney, F.P. & Black, J.H. 1988, ApJ, 325, 389.
16. Millar, T.J. 1997, in IAU Symp. #178, *Molecules in Astrophysics: Probes & Processes*, ed. E.F. van Dishoeck, Kluwer:Dordrecht, p75.
17. Millar, T.J., Bennett, A. & Herbst, E. 1989, ApJ, 340, 906.
18. Mochizuki, K. *et al.* 1994, ApJ, 430, L37.
19. Neufeld, D.A. & Dalgarno, A. 1989, ApJ, 340, 869.
20. Pineau des Forêts, G. & Flower, D. 1997, in IAU Symp. #178, *Molecules in Astrophysics: Probes & Processes*, ed. E.F. van Dishoeck, Kluwer:Dordrecht, p113.
21. Sternberg, A. 1990, ApJ, 361, 121.
22. Sternberg, A., Min Yan, & Dalgarno, A. 1997 in IAU Symp. #178, *Molecules in Astrophysics: Probes & Processes*, ed. E.F. van Dishoeck, Kluwer:Dordrecht, p141.
23. Taniguchi, Y. & Ohyama, Y. 1998, ApJ, 507, L121.
24. Tielens, A.G.G.M. & Hollenbach, D. 1985, ApJ, 291, 722.
25. van Dishoeck, E.F. & Black, J.H. 1988, ApJ, 334, 711.
26. Viala, Y.P., Letzelter, C., Eidelberg, M. & Rostas, F. 1988, A&A, 193, 265.
27. Williams, D.A. 1987, in *Physical Processes in Interstellar Clouds*, eds. G.E. Morfill & M. Scholer, Reidel:Dordrecht, p377.
28. Zhang, Q., Ho, P.T.P., Wright, M.C.H., & Wilner, D.J. 1995, ApJ, 451, L71.

14

1. Begelman, M.C. & McKee, C.F. 1990, ApJ, 358, 375.
2. Bronfman, L., Cohen, R.S., Alvarez, H., May, J. & Thaddeus, P. 1988, ApJ, 324, 248.
3. Boulanger, F. & Perault, M. 1988, ApJ, 330, 964.
4. Caulet, A. & Newell, R. 1996, ApJ, 465, 205.
5. Chièze, J-P. 1987, A&A, 171, 225.
6. Dame, T.M., Elmegreen, B.G., Cohen, R.S. & Thaddeus, P. 1986, ApJ, 305, 892.
7. Devereaux, N., Taniguchi, Y., Sanders, D.B., Nakai, N. & Young, J.S. 1994, AJ, 107, 2006.
8. Dopita, M.A., Mathewson, D.S., & Ford, V.L. 1985, ApJ, 297, 599.
9. Dopita, M.A., & Ryder, S.D.1994b, ApJ, 430, 163.
10. Dwek, E. & Sclo, J.M. 1980, ApJ, 239, 193.
11. Efremov, Y.N. & Elmegreen, B.G. 1998, MNRAS, 299, 643.
12. Egger, R.J. 1995, in *The Physics of the Interstellar & Intergalactic Medium*, eds. A. Ferrara, C.F. McKee, C. Heiles & P.R. Shapiro, ASP Conf. Ser. 80, 45.
13. Elmegreen, B.G. 1997, Rev. Mex. Ast. y Astrofis. Conf. Ser., 6, 165.
14. Elmegreen, B.G. 1998, in *Origins of Galaxies, Stars, Planets & Life*, eds. C.E. Woodward, H.A. Thronson & M. Shull, PASP Conf Ser. 148, p149.
15. Falgarone, E. 1995, in *The Physics of the Interstellar & Intergalactic Medium*, eds. A. Ferrara, C.F. McKee, C. Heiles & P.R. Shapiro, ASP Conf. Ser. 80, 158.
16. Falgarone, E. & Puget, J.P. 1986, A&A, 162, 235.
17. Ferrara, A. & Corbelli, E. 1995, in *The Physics of the Interstellar & Intergalactic Medium*, eds. A. Ferrara, C.F. McKee, C. Heiles & P.R. Shapiro, ASP Conf. Ser. 80, 346.
18. Ferriere, K.M., Zweibel, E.G. & Shull, J.M. 1988, ApJ, 332, 984.
19. Field, G.B. 1965, 142, 531.
20. Field, G.B., Goldsmith, D.W. & Habing, H.J. 1969, ApJ, 155, L149.
21. Fukui, Y. *et al.* 1988, ApJ, 311, L85.
22. Jenkins, E.B., 1987, in *Interstellar Processes*, eds. D.J. Hollenbach & H.A. Thronson, Jr., Reidel:Dordrecht, p533.
23. Jenkins, E.B., 1987, in *Interstellar Dust*, eds. L.J. Allamandola & A.G.G.M. Tielens, Reidel:Dordrecht, p23.
24. Jenkins, E.B., Jura, M. & Loewenstein, M. 1983, ApJ, 270, 88.
25. Kennicutt, R.C. Jr, 1983, ApJ, 272, 54.
26. Kennicutt, R.C. Jr, 1989, ApJ, 344, 685.
27. Kennicutt, R.C. Jr, 1998, ApJ, 498, 541.
28. Kennicutt, R.C., Bresolin, F., Bomans, D.J., Bothun, G.D., & Thompson, I.B., 1995, AJ, 109, 594.
29. Kim, S., *et al.* 1998a, ApJ, 503, 674.
30. Kim, S., *et al.* 1998b, PASA, 15, 132.

31. Kim, S., Chu, Y-H., Staveley-Smith, L. & Smith, R.C., ApJ, 503, 729.
32. Krolik, J.H., McKee, C.F. & Tarter, C.B. 1981, ApJ, 249, 422.
33. Larson, R.B. 1988, in *Galactic and Extragalactic Star Formation*, eds. R.E. Pudrich & M.Fitch, Kluwer: Dordrecht, p435.
34. Lada, C.J. 1988, in *Galactic and Extragalactic Star Formation*, eds. R.E. Pudritz & M. Fitch, Kluwer:Dordrecht, p5.
35. McCray, R. & Kafatos, M. 1987, ApJ, 317, 190.
36. McKee, C.F. 1989, ApJ, 345, 782.
37. McKee, C.F. 1989b, in *Interstellar Dust*, eds. L.J. Allamandola & A.G.G.M. Tielens, Reidel:Dordrecht, p431.
38. McKee, C.F. and Ostriker, J.P. 1977, ApJ, 218, 148.
39. Maddelena, R. & Thaddeus, P. 1985, ApJ, 294, 231.
40. Mathis, J.S., Mezger, P.G., & Panagia, N. 1983, A&A, 128, 212.
41. Meaburn, J. 1980, MNRAS, 289, 570.
42. Oey, M.S. & Clarke, C.J. 1997, MNRAS, 289, 570.
43. Oey, M.S. & Massey, P. 1995, ApJ, 452, 210.
44. Oey, M.S. & Smedley, S.A. 1998, AJ, 116, 1263.
45. Pagel, B. 1997, *Nucleosynthesis and the Chemical Evolution of Galaxies*, CUP:Cambridge.
46. Quirk, W.J. 1972, ApJ, 176, L9.
47. Raymond, J. 1992, ApJ, 384, 191.
48. Reynolds, R.J. 1980, ApJ, 236, 153.
49. Reynolds, R.J. 1984, ApJ, 282, 191.
50. Rosado, M. *et al.* 1996, A&A, 308, 588.
51. Ryder, S.D. & Dopita, M.A1994, ApJ, 430, 142.
52. Schmidt, M. 1959, ApJ, 129, 736.
53. Scoville, N.Z., Yun, M.S., Clemens, D.P., Sanders, D.B. & Waller, W.H. 1987, ApJS, 63, 821.
54. Silk, J. 1997, ApJ, 481, 702.
55. Sleath, J.P. & Alexander, P. 1995, MNRAS, 275, 507.
56. Solomon, P.M., Rivolo, A.R., Barrett, J.W. & Yahil, A. 1987, ApJ, 319, 730.
57. Stark, A.A. 1984, in *Kinematics, Dynamics & Structure of the Milky Way*, ed. W.L. Shuter, Reidel:Dordrecht, p127.
58. van der Kruit P.C. & Searle L. 1981a, A&A, 95, 105.
59. van der Kruit P.C. & Searle L. 1981b, A&A, 95, 116.
60. van der Kruit P.C. & Searle L. 1982, A&A, 110, 61.
61. van Dishoeck, E.F. & Black, J.H. 1986, ApJS, 64, 109.
62. Walborn, N.R. & Parker, J.W. 1992, ApJ, 399, L87.
63. Wang, Q. & Helfand, D.J. 1991, ApJ, 379, 327.
64. Wang, J., Heckman, T.M. & Lehnert, M.D. 1998, ApJ, 509, 93.
65. Weis, K., Chu, Y-H, Duschl, W.J. & Bomans, D.J. 1997, A&A, 325, 1157.
66. Wolfire, M.G., Hollenbach, D.J., McKee, C.F., Tielens, A.G.G.M. & Bakes, E.L.O. 1995, ApJ, 443, 152.
67. Wright, E.L.*et al.* 1991, ApJ, 381, 200.
68. Wyse, R.F.G. 1986, ApJ, 311, L41.