Radiative Feedback and the First Stars

Simon Glover

Doctor of Philosophy
The University of Edinburgh
2001

This thesis is my own composition except where indicated in the text.

July 18, 2002

Acknowledgments

First and foremost, I would like to express my gratitude to my supervisor, Peter Brand, for suggesting such an interesting project to work on in the first place, for providing patient guidance over the past few years, and for asking difficult questions at the right time.

I would also like to thank all of the many people with whom I've discussed aspects of this work, or whose suggestions have helped to improve it: Tom Abel, Omar Almaini, Rennan Barkana, John Black, Bruce Draine, Marie Machacek, Avery Meiksin, Louisa Nolan, Si Peng Oh, John Peacock, Susie Scott and Rob Smith.

My thanks also go to all of the people that helped make Edinburgh such a nice place to live and work: Alison, Ali, Alex, Andy D, Andy T, Ben, Chris, Dan, Dave, Eelco, Elese, Helen, Ian, James, Jo, Kenny, Licia, Lou, Marek, Matt, Meghan, Michael, Olivia, Omar, Pete, Phil, Rachel, Richard B & Richard K, Rob, Ross, Simon, Sujata, Susie, Tara, Thomas, Will P & Will S. It wouldn't have been nearly so much fun without you.

Finally, my deepest personal thanks go to Louisa (for being there when I needed a friend), Susie (for being there when I needed a drink), and Alex & Christiane (just for being).

Abstract

Numerical simulations suggest that the very first stars to form do so within cool gas in small protogalaxies. These protogalaxies have low virial temperatures, and cooling within them is dominated by molecular hydrogen, H_2. This is easily destroyed by ultraviolet radiation from newly-formed stars, and this 'radiative feedback' may play an important role in regulating star formation in the early universe.

To study the effects of radiative feedback, requires an accurate chemical model. Examples currently in the literature assume the gas to be optically thin, and give erroneous results in optically thick gas. In chapter 2 of this thesis I develop a chemical model that correctly treats the photochemistry of optically thick primordial gas. I also discuss the approximations that remain, and estimate the accuracy of the model.

In chapter 3, I examine the role of radiative feedback on small scales. Using a simple protogalactic model, I determine the growth timescales and final sizes of H II regions within a protogalaxy and discuss the effect of photoionizing radiation on dense clumps of gas. I also examine the effects of photodissociation, and present a simple method for estimating the photodissociation timescale in optically thick gas. I find that radiative feedback occurs rapidly in diffuse protogalactic gas, but that dense clumps can resist its effects and survive to form stars.

On larger scales, previous work has shown that early star formation rapidly builds up a soft UV background that can suppress star formation within the smallest protogalaxies. However, this work does not include the effects of X-rays produced by sources associated with star formation, which will catalyze H_2 formation and reduce the effects of photodissociation. In chapter 4, I examine the importance of these X-rays by self-consistently modelling the growth of the X-ray and UV

backgrounds together with their effects on gas within protogalaxies. An important result is the determination of $T_{\rm crit}$, the critical virial temperature above which protogalaxies can cool. The evolution of $T_{\rm crit}$ is presented for various X-ray source models, and compared to a model with no X-ray background, allowing the effects of X-ray feedback to be assessed.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

How does a star form? This seems a simple enough question and yet, despite decades of research by many, many talented people, there is still much that we do not know. In principle at least, forming a star is very easy – we simply need to take diffuse gas, for example from the interstellar medium, and increase its density by around twenty-five orders of magnitude until it becomes sufficiently dense and hot for nuclear fusion to begin. This compression could be initiated in any number of ways, but eventually we must rely on the self-gravity of the gas to drive the process. So far so good, but we soon find that a great many processes operate to oppose this collapse – for example, rotation, magnetic fields, thermal pressure and turbulence to name but a few. Understanding the effects of these processes, especially when several are operating at once, is one of the ongoing tasks within the present-day study of star formation.

In this thesis, I study one small part of this huge subject – the effect of ultraviolet and X-ray radiation on pre-stellar gas [1] at high redshift. In particular, I study something known as *radiative feedback* – the way in which radiation from newly-formed stars can promote or (more usually) suppress further star formation. Before I can properly begin to discuss the details of radiative feedback, however, it is necessary to place the discussion in its proper cosmological context. Accordingly, I begin this introduction with the discussion of a number of cosmological preliminaries; readers familiar with this material may wish to skip ahead to section 1.3 or 1.4.

[1] Gas from which we expect stars to form.

1.1 Cosmological Preliminaries

1.1.1 The Cosmological 'Standard Model'

Following Peebles (1993) in borrowing a concept from particle physics, we can identify a 'standard model' of physical cosmology – a world picture which, although incomplete, has survived many observational tests and is consistent with the available evidence.

This model asserts that the distribution of matter in the universe is homogeneous and isotropic if smoothed on a sufficiently large scale, that it is expanding and that it was substantially hotter and denser in the past. It is based upon two assumptions and a number of important observations.

The first assumption is simply that the laws of physics on cosmological scales are the same as those on smaller scales; in particular, that general relativity is the correct description of gravity on large scales. This assumption is certainly consistent with the evidence, but it is wise to remember that it involves an extrapolation from the terrestrial scale on which GR is well tested to a scale that is many orders of magnitude larger.

The other assumption that we make is often termed the Copernican Principle: the assumption that there is nothing particularly special about the position we occupy in the universe. How true this is on a sub-galactic scale is still open to question – as far as we know, the existence of life requires a number of special conditions (such as the existence of liquid water), which are not ubiquitous. On the other hand, there appears to be nothing particularly special about the Milky Way compared with, say, M31 or M101, and there seems to be no good reason to suppose that our presence in this particular galaxy is due to anything other than chance.

Turning to the observations, we find a mixture of direct and indirect evidence for the model. The cosmological microwave background (CMB) and the hard X-ray background both provide strong evidence of isotropy (Wu *et al.*, 1999). Moreover, this also is strong evidence for homogeneity; the observed isotropy is a natural consequence of a homogeneous universe, but would require us to be in a particularly special location in an inhomogeneous universe. We can also test homogeneity directly, by means of galaxy counts. On small scales, galaxies are

known to cluster inhomogeneously, but if we smooth the number counts on larger and larger scales, we expect to see a transition to homogeneity. There is some controversy over whether or not this transition has yet been detected (di Nella *et al.*, 1996; Hatton, 1999), but the question should be settled by the various large redshift surveys currently underway, such as 2dF (Maddox, 2000) or the Sloan Digital Sky Survey (York *et al.*, 2000).

Evidence for expansion comes from the redshift-distance relation: the fact that the spectra of extragalactic sources are shifted to the red by an amount that is directly proportional to distance. This was first noted by Wirtz (1924), but was first quantified by Hubble (1929), and has since become known as Hubble's law. It is a natural consequence of an expanding universe, so long as the expansion is homogeneous.

Another natural consequence of an expanding universe, as long as energy is conserved, is the fact that it must have been hotter and denser in the past. Evidence that this was indeed the case comes from two different sources: the observed elemental abundances and the microwave background.

By observing elemental abundances in regions that have undergone little or no star formation (Adams, 1976), we can attempt to infer the primordial abundances of the various elements. We find that the inferred abundances are consistent with the predictions of the expanding universe model, to within the observational uncertainties (Burles *et al.*, 2000a). Moreover, the theoretical model of primordial nucleosynthesis has only one free parameter – the baryon-photon ratio – so this success is impressive.

The microwave background is also strong evidence for a hot,dense phase. In particular, measurements by the FIRAS experiment on the COBE satellite have shown that the CMB has a spectrum which is an almost perfect black-body; any deviation is at the level of one part in ten thousand or less (Fixsen *et al.*, 1996). This observation is simple to explain in the standard model. At an early epoch, the temperature is high enough that all of the gas is ionized and consequently radiation and matter are closely coupled by Thomson scattering and in thermal equilibrium. At this time, the radiation spectrum is necessarily that of a black-body. As the universe expands, however, the temperature drops, and the gas eventually begins to recombine. At this point, the Thomson scattering optical depth drops sharply. The bulk of the radiation never subsequently interacts

with matter, and thus, aside from the effects of redshift, [2] the spectrum remains unchanged to the present day.

To produce a black-body CMB from other sources of radiation would be far harder. It would require the action of some mechanism capable of thermalizing the spectrum, but at the same time having no discernable effects on the spectra of extragalactic sources visible at the present day.

Individually, none of these pieces of evidence are entirely persuasive – if we tried hard enough we could usually construct *some* model to explain them. Taken together, however, the fact that the *same* simple model explains all of the otherwise unrelated observational evidence argues strongly for its basic correctness.

The standard model by no means represents all of cosmology; there are many areas, such as the formation of galaxies, or the evolution of the universe prior to nucleosynthesis, on which it stands mute. Nevertheless, it provides us with a basic framework within which our other theories must fit. In the next section, I examine some of the details of this framework.

1.1.2 Basic Equations

The Robertson-Walker Metric

On the largest scales our universe is dominated by gravity and it is to general relativity that we must turn first in our attempt to understand it. Our starting point is the large-scale structure of spacetime itself, as encapsulated in the form of the space-time metric. This can be written in its most general form as (Peacock, 1999)

$$c^2 \, \mathrm{d}\tau^2 = g_{\mu\nu} \, \mathrm{d}x^\mu \, \mathrm{d}x^\nu, \tag{1.1}$$

where $\mathrm{d}\tau$ is a proper time interval and the $\mathrm{d}x^\mu$ are coordinate intervals; we follow the convention that Greek indices run from zero to three, with x^0 being the time coordinate and x^1, x^2, x^3 the spatial coordinates. The metric tensor $g_{\mu\nu}$ is in principle a function of all four coordinates and is usually obtained by solving the GR field equations

$$G^{\mu\nu} = -\frac{8\pi G}{c^4} T^{\mu\nu}. \tag{1.2}$$

[2] Redshifting a black-body spectrum does not alter its shape, but merely lowers its characteristic temperature

for every point in spacetime. However, if the distribution of matter is homogeneous and isotropic then this places significant constraints on the form of $g_{\mu\nu}$ and allows much of its form to be deduced purely on the basis of symmetry arguments.

We begin by noting that in an expanding, homogeneous universe, it is always possible for observers who are at rest with respect to the matter distribution (also known as fundamental observers) to agree on the definition of a universal time. For instance, they could agree to synchronize clocks when the density of the universe reached some predetermined value – following this, homogeneity implies that the clocks must remain synchronized.

Next, consider two events separated by a small distance in space-time. A local observer can always choose coordinates such that the separation is given by

$$c^2\,\mathrm{d}\tau^2 = c^2\,\mathrm{d}t^2 - \mathrm{d}x^2 - \mathrm{d}y^2 - \mathrm{d}z^2. \tag{1.3}$$

But we know that the local observer's time coordinate is the same as that of a distant observer, so the distant observer will see this separation as

$$c^2\,\mathrm{d}\tau^2 = c^2\,\mathrm{d}t^2 - g_{ij}\,\mathrm{d}x^i\,\mathrm{d}x^j \tag{1.4}$$

where the Latin indices run from one to three.

Isotropy allows another simplification. If we write the metric in terms of the spherical polar coordinates (r, θ, ϕ), then isotropy implies that g_{ij} must be independent of θ and ϕ. Moreover, homogeneity implies that the time dependence of g_{ij} (if any) can be separated from its spatial dependence. Together, these simplifications allow us to write the metric as

$$c^2\,\mathrm{d}\tau^2 = c^2\,\mathrm{d}t^2 - R(t)^2[f(r)\,\mathrm{d}r^2 + g(r)\,\mathrm{d}\Omega^2] \tag{1.5}$$

where $\mathrm{d}\Omega^2 = \mathrm{d}\theta^2 + \sin^2\theta\,\mathrm{d}\phi^2$. The metric is fully specified by a time dependent scale factor $R(t)$ and two functions of radial position, f and g.

As yet r is simply some way of labelling points in the radial direction and, aside from a few constraints such as continuity and monotonicity, it is largely arbitrary. Indeed, if we give $R(t)$ dimensions of length, then it is not even a real length; rather, it is simply a dimensionless number. This gives us the freedom to choose a definition for r that makes life simple, either by ensuring that $f = 1$ or that $g = r^2$.

The form of the remaining spatial function is then fixed by the curvature of space. Homogeneity and isotropy again force substantial simplification – space can be

positively curved, negatively curved or flat, but the curvature must be the same everywhere. This leads to the following general form of the metric

$$c^2\,\mathrm{d}\tau^2 = c^2\,\mathrm{d}t^2 - R(t)^2[\,\mathrm{d}r^2 + S_k^2(r)\,\mathrm{d}\Omega^2], \tag{1.6}$$

where

$$S_k(r) = \begin{cases} \sin r & k = 1 \\ \sinh r & k = -1 \\ r & k = 0. \end{cases} \tag{1.7}$$

and k is a convenient way to parameterise the curvature – positive k implies positive curvature, negative k implies negative curvature while $k = 0$ implies flat space.

That this is the most general form of the metric in a homogeneous and isotropic universe was first demonstrated, independently, by Robertson (1935) and Walker (1936); it is generally called the Robertson-Walker metric.

Redshift

We can deduce a number of properties of the universe directly from the form of the metric. A particularly important example is the apparent time dilation of distant events – this also implies the redshifting of light from distant objects.

Consider a distant galaxy, at a point r. Light from this galaxy reaches us along a null geodesic, so

$$c^2\,\mathrm{d}t^2 = R^2\,\mathrm{d}r^2, \tag{1.8}$$

and a simple integration gives the comoving distance

$$r = \int \frac{c\,\mathrm{d}t}{R}. \tag{1.9}$$

Now consider a pulse of light emitted from this galaxy at a time t_e with a duration δt_e, which is observed at a time t_o to have a duration δt_o. The comoving distance travelled by the start of this pulse is

$$\int_{t_e}^{t_o} \frac{c\,\mathrm{d}t}{R}, \tag{1.10}$$

while the end travels a distance

$$\int_{t_e+\delta t_e}^{t_o+\delta t_o} \frac{c\,\mathrm{d}t}{R}. \tag{1.11}$$

But r has remained constant, so these two distances must be the same. If the length of the pulse is small, this implies that

$$\frac{\delta t_o}{R(t_o)} = \frac{\delta t_e}{R(t_e)} \tag{1.12}$$

and since $R(t_o) > R(t_e)$ in an expanding universe, we observe a time dilation. If, instead of a pulse of light, we apply the same argument to a monochromatic light beam, we see that it will be redshifted by an amount

$$1 + z = \frac{R(t_o)}{R(t_e)}. \tag{1.13}$$

The Friedmann equations

In order to make use of the metric, we need to know how $R(t)$ evolves. Solution of the field equations gives us two equations for R:

$$\dot{R}^2 - \frac{8\pi G}{3}\rho R^2 = -kc^2 \tag{1.14}$$

and

$$\frac{\ddot{R}}{R} = -\frac{4\pi}{3}G\left(\rho + \frac{3p}{c^2}\right) \tag{1.15}$$

where ρ is the total energy density and p the corresponding pressure.

A spatially flat universe requires that $k = 0$. This is only possible if $\rho = \rho_{\rm crit}$, where

$$\rho_{\rm crit} = \frac{3H^2}{8\pi G} \tag{1.16}$$

and $H \equiv \dot{R}/R$ is the Hubble parameter [3] This can be more conveniently expressed in terms of the dimensionless density parameter Ω

$$\Omega = \frac{\rho}{\rho_{\rm crit}}. \tag{1.17}$$

A value of $\Omega = 1$ corresponds to a flat universe, while $\Omega > 1$ implies a closed universe and $\Omega < 1$ an open universe. In general, the value of Ω is time-dependent; I denote its value at a redshift z as $\Omega(z)$, and the value at the present day as Ω.

To solve equations 1.14 and 1.15 for R, we need to know $\dot{\rho}$. This is obtained from the principle of conservation of energy – we can write the familiar equation

$$\mathrm{d}U = -p\,\mathrm{d}V \tag{1.18}$$

[3] Its value at the present day is denoted H_0 and is known as the Hubble constant.

in terms of the energy density and scale factor as

$$\mathrm{d}(\rho c^2 R^3) = -p\,\mathrm{d}(R^3), \tag{1.19}$$

implying that

$$\dot{\rho} = -3H(z)\left(\rho + \frac{p}{c^2}\right). \tag{1.20}$$

Equations 1.14 and 1.15, together with the above equation, are commonly known as the Friedmann equations (Friedmann, 1922, 1924). To solve them, we need only the equation of state. This may be quite complex as ρ consists of contributions from a number of different sources, each with its own equation of state. Fortunately, however, these contributions are additive, and we can consider the possible components individually.

Dust

In this context, dust refers to matter with no pressure, *i.e.* for which $p = 0$. This is clearly an idealization, but is a good approximation when $p \ll \rho c^2$, which is true for matter with non-relativistic thermal velocities. With this approximation, equation 1.20 becomes

$$\dot{\rho}_m = -3H(z)\rho_m \tag{1.21}$$

with solution

$$\rho_m = \rho_0(1+z)^3, \tag{1.22}$$

where ρ_0 is the value at the present day.

Radiation

Radiation has $p = \frac{1}{3}\rho c^2$; this is also an accurate approximation for matter with relativistic thermal velocities. In this case, equation 1.20 becomes

$$\dot{\rho}_r = -4H(z)\rho_r \tag{1.23}$$

with solution

$$\rho_r = \rho_{r0}(1+z)^4. \tag{1.24}$$

Vacuum energy

A third possible contribution comes from something which is generally known as vacuum energy. The basic idea is that any energy density associated with empty space will have a negative equation of state. To see why this is so, consider the effects of expanding space by an infinitesimal amount $\mathrm{d}V$. This increases the

energy content by $\mathrm{d}U = \rho_v c^2 \,\mathrm{d}V$, and since we know that $\mathrm{d}U = -p\,\mathrm{d}V$ for an adiabatic change, the pressure must be given by

$$p_v = -\rho_v c^2. \tag{1.25}$$

This further implies that

$$\dot{\rho}_v = 0, \tag{1.26}$$

i.e. the vacuum energy density remains constant. This has the same effect within the field equations as the addition of a constant term $\Lambda g^{\mu\nu}$ to the left hand side of equation 1.2, provided that $\Lambda = 8\pi G\rho_v/c^2$. A contribution of this form, termed the cosmological constant, was first suggested by Einstein in order to allow for static solutions of the field equations. Although he later retracted his suggestion, in the light of Hubble's demonstration that the universe was expanding, the possibility of a contribution from a cosmological constant has been debated ever since. The idea that a cosmological constant could be explained in terms of a vacuum energy contribution was first introduced by Zeldovich (1967).

How seriously should we take this idea? After all, how can empty space contain energy? The answer to this question is not particularly well understood. Quantum theory gives us some insight. In particular, it shows us that there is no such thing as truly empty space. Rather, the vacuum state in quantum field theory is simply the minimum energy state. This energy of this minimum *may* be zero, but there seems to be nothing in quantum theory which forces it to be so. If, for some reason, it is non-zero, then it would contribute to the energy-momentum tensor as a vacuum energy density.

There is, however, a serious problem. The only natural values for ρ_v to have, other than zero, are set by the energy scales of the various phase transitions that the fields undergo. These are orders of magnitude too large, vastly exceeding the observational bound on ρ_v. Some possible solutions to this problem are examined in Weinberg (1989), but it remains true to say that nobody really understands why ρ_v should be anything other than zero.

So why consider it at all? Well, recent observations seem to suggest that we can't do without it; they are inconsistent with cosmological models in which $\Lambda = 0$. This conclusion rests on two different types of observation. Firstly, there are observations of type Ia supernovae at high redshift which have been made by two different groups (Perlmutter *et al.*, 1999; Riess *et al.*, 1998). Their aim is to

determine the deceleration parameter q_0, which is defined as

$$q_0 \equiv -\frac{\ddot{R}R}{\dot{R}^2} \tag{1.27}$$

$$= \frac{\Omega_m}{2} + \Omega_r - \Omega_v, \tag{1.28}$$

where

$$\Omega_i \equiv \frac{\rho_i}{\rho_{crit}}. \tag{1.29}$$

(Note that Ω_r is negligible at the present day). The observations strongly suggest that $q_0 < 0$; in other words, that the universe is accelerating. If true, this requires $\Omega_v > 0$; however, the observations are also marginally consistent with models in which $\Lambda = 0$ and $\Omega \lesssim 0.4$.

On the other hand, observations of the angular power spectrum of the anisotropies in the CMB can also be used to constrain Ω. Recent balloon-based measurements (De Bernadis *et al.*, 2000; Hanany *et al.*, 2000) provide strong evidence for a flat (or nearly flat) universe: $\Omega = 1.07 \pm 0.06$.

Taken together, these two observations are strong evidence for non-zero Λ; both sets of data can be well fit by a cosmological model with $\Omega_m = 0.3$ and $\Omega_v = 0.7$. Such a model is also consistent with local dynamical estimates of Ω_m, provided that galaxies are reasonably unbiased tracers of the mass distribution (Valentine *et al.*, 2000).

This is not a particularly comfortable conclusion, as it suggests that we simply do not understand what the bulk of the energy density of the universe actually *is*; however, unless the observations are seriously in error, the conclusion seems to be unavoidable.

Returning to our main theme, let us suppose that we know the current values of ρ_m, ρ_r and ρ_v. In this case, it is straightforward to solve equation 1.14 for the evolution of the scale factor. We can write the Hubble parameter as

$$H(z)^2 = H_0^2 \left[\Omega_v + \Omega_m(1+z)^3 + \Omega_r(1+z)^4 + (1-\Omega)(1+z)^2\right]. \tag{1.30}$$

When $\Omega_v = 0$, this has two types of solution: for $\Omega > 1$ the universe expands and then recollapses, while for $\Omega \leq 1$ expansion continues indefinitely. For $\Omega_v \neq 0$, the solutions are more complicated. They are analyzed in detail in Carroll *et al.* (1992).

For the special case of $\Omega = \Omega_m = 1$, equation 1.30 takes the particularly simple form

$$H(z) = H_0(1+z)^{3/2}. \tag{1.31}$$

In this case, it is simple to solve for the cosmological time

$$t = \frac{t_0}{(1+z)^{3/2}}, \tag{1.32}$$

where t_0 is the current age of the universe, given by

$$t_0 = \frac{2}{3H_0}. \tag{1.33}$$

This model is commonly known as the Einstein-de Sitter model. Although it is ruled out by the supernova observations, it is still of interest due both to its simplicity and also to the fact that it becomes an increasingly good approximation to the general solution as we go to higher redshift.

A large number of other useful results can be extracted from the Friedmann equations and the Robertson-Walker metric; some are derived in later chapters, while for the rest, I refer the reader to any basic cosmology textbook.

1.1.3 Early Epochs

The earliest stages of the universe involve many complicated physical processes (Kolb & Turner, 1990) occuring at very high temperatures (and hence very high energies). Many of these occur at or beyond the limits of our current understanding of particle physics [4] and there remains much about this period that is unclear. Fortunately, little of this complexity is relevant to the study of galaxy formation: recombination and nucleosynthesis give us a good idea of what the universe was like at somewhat lower redshift, and the question of how it got to be that way is rather less important. Nevertheless, one thing that is believed to occur at this time *does* have important consequences – the period known as inflation.

Inflation

It is widely believed amongst cosmologists that the universe must have passed through a phase of strongly accelerated expansion or *inflation*. The general idea

[4] Or, at the very least, at the limits of that part of particle physics that we have been able to experimentally test.

is that in the early universe there exists some scalar field (called an *inflaton*) with a non-zero potential. This potential is a source of vacuum energy and will act as a large cosmological constant. If sufficiently large, it will completely dominate the expansion, in which case the scale factor will evolve as (Peacock, 1999)

$$R(t) \propto \begin{cases} \cosh Ht & k = 1 \\ \sinh Ht & k = -1 \\ \exp Ht & k = 0, \end{cases} \tag{1.34}$$

where $H = \sqrt{8\pi G\rho_v/3}$; hence, we get exponential expansion. Clearly, the field cannot be static; rather, we suppose that it evolves quickly towards the true zero of the potential. Upon reaching the zero point, the remaining energy density of the field is converted to matter and radiation, through couplings with the matter fields, in a process called reheating (Linde, 1996). We are thus left with a conventional universe containing matter and radiation that subsequently evolves in a more familiar manner.

The identity of the field that drives inflation is not known – the original suggestion was that it was the Higgs field (Guth, 1981), but it was soon realised that this would not work (Guth & Weinberg, 1983). A number of other possibilities have been suggested (see *e.g.* Liddle, 1997, and references therein) but there is little consensus. Given this, and the fact that experimental particle physics provides no real motivation for the existence of an inflaton, why is the idea of inflation taken seriously? The answer is that inflation solves a number of long-standing cosmological problems.

Firstly, it explains the observed flatness of the universe (*i.e.* the fact that $\Omega \sim 1$) without the need for early fine-tuning; exponential expansion leaves the universe with

$$\Omega = 1 + O(f^{-2}) \tag{1.35}$$

where f is the number of e-foldings of the expansion.

It also explains the large-scale homogeneity of the universe. This is difficult to explain in the standard model, which predicts that the particle horizon (Peacock, 1999)

$$r_H = \int_0^t \frac{c\,\mathrm{d}t}{R(t)} \tag{1.36}$$

has a size at last-scattering of the order of 100 Mpc, corresponding to about a degree on the sky. This implies that the observed microwave background consists

of light from many different causally disconnected regions, and yet all of these regions contrive to have exactly the same temperature (at least to within the μK scale of the observed anisotropies). Inflation provides a simple solution to this problem: all of the regions that we see were initially causally connected, but became disconnected during the period of exponential expansion.

A third problem solved by inflation, indeed the one which motivated its original invention, is the monopole problem. Grand unified theory (GUT) models predict that a large number density of monopoles – point-like topological defects – should be produced during the breaking of the GUT symmetry in the early universe. However, the predicted number density is orders of magnitude greater than the observational upper limit. Inflation again provides a natural solution to this problem – if monopoles are generated before the period of exponential expansion, then inflation rapidly reduces their density to almost zero.

Finally, inflation provides a natural method for producing large scale density perturbations. On very small scales, the uncertainty principle implies the existence of small fluctuations. During a period of exponential expansion, these fluctuations will also expand, and, if the expansion is fast enough then they will be reach super-horizon scales and become 'frozen-in' before they can die away. As a result, inflation generates small perturbations to the background over a very wide range of scales; in particular, one can obtain perturbations on scales that are larger than the particle horizon, which would be not be possible in any other model. As well as straightforward density perturbations, gravitational waves can also be generated on a range of scales by this mechanism. Detection of these waves by their imprint on the CMB (Lesgourgues *et al.*, 2000) presents the most immediate prospect for an observational test of inflation.

1.1.4 Nucleosynthesis

The next event of direct relevance to galaxy formation occurs at $z \sim 10^{10}$, when primordial baryons begin to fuse together to form heavier elements. This occurs in several stages.

At very high temperatures, one has only protons and neutrons, with abundances fixed by thermal equilibrium

$$\frac{n_n}{n_p} = \exp\left(\frac{-\Delta mc^2}{kT}\right) \tag{1.37}$$

$$\simeq \exp\left(\frac{-1.5 \times 10^{10}\,\mathrm{K}}{T}\right) \tag{1.38}$$

Any heavier nuclei which form are rapidly photodissociated. At very high temperatures, $n_n \simeq n_p$, but as the universe expands and the temperature falls, the balance begins to shift in favour of protons. If nothing else were to intervene, then all of the neutrons would rapidly be converted to protons. However, at a temperature of roughly 10^{10} K, the weak interaction timescale becomes longer than the expansion timescale and the neutron-proton ratio freezes out at approximately 0.34.

After freeze-out, beta decay begins to reduce the neutron abundance, on a timescale of $t = 887$ s. However, before it can have a significant impact, the temperature drops enough that it becomes possible to form significant quantities of heavier nuclei. Consequently, most of the neutrons do not decay, but instead fuse with protons to form deuterium. This in turn fuses to ^{3}He via the reactions

$$\mathrm{D} + \mathrm{p} \rightarrow {}^3\mathrm{He} + \gamma \tag{1.39}$$

$$\mathrm{D} + \mathrm{D} \rightarrow {}^3\mathrm{He} + \mathrm{n}, \tag{1.40}$$

or to ^{3}H via

$$\mathrm{D} + \mathrm{n} \rightarrow {}^3\mathrm{H} + \gamma \tag{1.41}$$

$$\mathrm{D} + \mathrm{D} \rightarrow {}^3\mathrm{H} + \mathrm{p} \tag{1.42}$$

(Salati, 1997). ^{4}He is subsequently formed via

$${}^3\mathrm{He} + \mathrm{D} \rightarrow {}^4\mathrm{He} + \mathrm{p} \tag{1.43}$$

$${}^3\mathrm{H} + \mathrm{D} \rightarrow {}^4\mathrm{He} + \mathrm{n}. \tag{1.44}$$

The binding energy per nucleon of ^{4}He is substantially larger than that of deuterium, tritium or ^{3}He, so any ^{4}He that forms is unlikely to be broken down. Consequently, almost all of the neutrons become locked up in helium; the abundances of the intermediate species at the end of nucleosynthesis are small.

It would be reasonable to assume that ^{4}He would in turn fuse to form heavier elements. However, aside from a small amount of ^{7}Li and ^{7}Be, no elements heavier than ^{4}He form. This is because the obvious reactions

$$\mathrm{p} + {}^4\mathrm{He} \rightarrow {}^5\mathrm{Li} \tag{1.45}$$

$${}^4\mathrm{He} + {}^4\mathrm{He} \rightarrow {}^8\mathrm{Be} \tag{1.46}$$

produce isotopes which are extremely unstable and rapidly decay back to their initial components, while the triple-α reaction

$$^4\mathrm{He} + ^4\mathrm{He} + ^4\mathrm{He} \rightarrow ^{12}\mathrm{C} + \gamma, \tag{1.47}$$

responsible for the first stages of heavy element synthesis in stars, proceeds at a negligible rate.

We are thus left with a gas which consists of roughly three parts, by mass, of hydrogen to one part of helium, with trace abundances of deuterium, ^{3}He and ^{7}Li. The precise abundances of these nuclei are controlled by a single free parameter, the baryon to photon ratio, η. Moreover, since the photon abundance at nucleosynthesis can be deduced from the temperature of the microwave background at the present day, knowledge of η implies knowledge of the baryon density and vice versa. Specifically, η is related to the combination $\Omega_b h^2$ by

$$\eta = 2.746 \times 10^{-8} \left(\frac{2.728\ \mathrm{K}}{T_r} \right)^3 \Omega_b h^2, \tag{1.48}$$

where T_r is the temperature of the CMB, Ω_b is the contribution to Ω from baryonic matter and h is the Hubble constant in units of $100 \mathrm{kms}^{-1}\mathrm{Mpc}^{-1}$. A measurement of $\Omega_b h^2$ would thus suffice to determine the primordial abundances. In practice, however, $\Omega_b h^2$ is difficult to determine directly and we usually do the reverse – use measurements of primordial abundances to infer η and hence $\Omega_b h^2$.

Determination of the primordial abundance of any of the elements is substantially complicated by the fact that most of the gas which we observe will have been contaminated by elements produced by stellar nucleosynthesis. The best technique to date for avoiding this problem was originally suggested by Adams (1976) and involves determining the D/H ratio in absorption line systems observed in the spectra of high redshift quasars. Deuterium is not produced during stellar nucleosynthesis (although it can be destroyed) and the observed D/H ratio is thus a strong lower limit on the primordial ratio. Moreover, the observed metal abundance of the absorption line systems allows one to estimate the fraction of the gas that has been processed through stars, and thus the fraction of primordial deuterium that has been lost. Typical absorption line systems have $0.001 \leq Z/Z_\odot \leq 0.01$, corresponding to a loss of about 1% of the primordial deuterium.

Recent measurements of the D/H ratio from absorption systems in the spectra of three quasars (see O'Meara *et al.*, 2001, and references therein) give a value of

Element		Abundance (relative to hydrogen)
Hydrogen (^{1}H)		1.0
Deuterium (^{2}H)		$3.0 \pm 0.4 \times 10^{-5}$
Helium	(^{3}He)	$1.09 \pm 0.06 \times 10^{-5}$
	(^{4}He)	0.0826 ± 0.0003
Lithium (^{7}Li)		$3.8^{+1.0}_{-0.8} \times 10^{-10}$

Table 1.1: The abundance by number, relative to hydrogen, of the elements formed during primordial nucleosynthesis. The value of D/H is based on the observational measurements of O'Meara *et al.* (2001); the other values are theoretical predictions based on this deuterium abundance and are taken from the same paper.

$D/H = 3.0 \pm 0.4 \times 10^{-5}$, corresponding to a baryon density $\Omega_b h^2 = 0.0205 \pm 0.0009$. The resulting primordial abundances are given in table 1.1, and are generally consistent with direct observational determinations (Tytler *et al.*, 2000; Burles *et al.*, 2000b; Olive *et al.*, 2000), particularly if systematic uncertainties remain in the measurements of the helium abundance (Pagel, 2000). Since most current determinations agree that $h = 0.65 \pm 0.15$, the nucleosynthesis results imply that $\Omega_b < 0.1$ – the baryon density falls short of the closure density by an order of magnitude. On the other hand, CMB observations imply that $\Omega \simeq 1$. There is a growing consensus that much of the discrepancy is made up by some form of vacuum energy, although whether in the form of a classical cosmological constant or a more exotic solution such as quintessence (Caldwell *et al.*, 1998) is not yet known. In addition to this, there is also a great deal of evidence for a third contribution, from a different form of matter. Specifically, observations of galaxy clusters suggest that only around 10 – 20% of the mass is accounted for by the visible gas and stars, with the remainder being made up of some dark component, commonly known as dark matter.

Although we have yet to detect dark matter directly, it is nevertheless an appealing hypothesis as it solves a number of astrophysical problems. For instance, the rotation curves of most galaxies remain flat in their outer regions, rather than declining in the way that we would expect if mass traced light. This is easily ex-

plained by supposing that galaxies sit in extended dark matter haloes. Moreover, as discussed in section 1.2, dark matter provides a way for perturbations to grow on small scales prior to matter-radiation decoupling, with the consequence that dark matter models can explain the observed large-scale structure of the universe in a way that purely baryonic models cannot. Various other lines of evidence for the existence of dark matter can be advanced (see, for example, the discussion in section 12.3 of Peacock, 1999) and overall the case is a strong one, despite the fact that its nature remains a matter of speculation.

1.1.5 Recombination

At a much lower redshift ($z \sim 1000$), the temperature and the density of this plasma fall sufficiently for it to begin to recombine. A rigorous treatment of recombination is possible (Seager *et al.*, 2000), but highly involved, while a highly simplified treatment (Peebles, 1993) gives surprisingly accurate results.

We begin by ignoring the complexities of helium recombination and assuming that it has completely recombined by the time that hydrogen begins to recombine. This is not strictly accurate (Seager *et al.*, 2000) but is an adequate approximation, particularly as helium contributes only 10% of the total ionization. We further assume that we can model hydrogen accurately by considering only three electronic states – the 1s ground state, plus the 2s and 2p excited states. With these approximations, we can write the net rate of production of hydrogen atoms as

$$\frac{\mathrm{d}n_{\mathrm{H}}}{\mathrm{d}t} = k_{2b} n_e^2 - \beta_e n_{2s} \tag{1.49}$$

$$= R + \Lambda \left(n_{2s} - n_{1s} e^{-h\nu_\alpha / kT} \right) . \tag{1.50}$$

The first term on the right-hand side of equation 1.49 is the usual case B recombination rate, incorporating recombination to all of the excited states but not the ground state. The second term represents ionization from all of these excited states. The ionization rate is proportional to the population of the 2s state because interactions with the thermal background radiation will rapidly bring the relative populations of the excited states into thermal equilibrium. The constant of proportionality, β_e, can be obtained using the principle of detailed balance. We know that, in thermal equilibrium, ionization from and recombination to excited

states must balance:

$$k_{2b} n_e^2 = \beta_e n_{2s}, \tag{1.51}$$

while the Saha equation tells us that

$$\frac{n_e^2}{n_{2s}} = \frac{(2\pi m_e kT)^{3/2}}{h^3} e^{-I_2/kT}, \tag{1.52}$$

where $I_2 = 3.4$ eV is the binding energy of the $n = 2$ energy level. Combining these equations, we find that

$$\beta_e = k_{2b} \frac{(2\pi m_e kT)^{3/2}}{h^3} e^{-I_2/kT}. \tag{1.53}$$

Now, although we have derived this relationship by assuming thermal equilibrium, it depends only upon atomic properties (plus the temperature) and remains true even when equation 1.51 is not satisfied.

Returning to equation 1.49, note that the difference between the recombination and ionization terms must correspond to the net rate of recombinations to the ground state. This can be written (as in equation 1.50) as the sum of the net recombinations rates from the 2s and 2p states.

Recombination from the 2s state to the ground state is strongly forbidden, and proceeds only via two-photon emission, with rate coefficient (Goldman, 1989)

$$\Lambda = 8.22458 \text{ s}^{-1}. \tag{1.54}$$

The rate of the inverse process can again be derived using detailed balance; it is $\Lambda\, e^{-h\nu_\alpha/kT}$, where $h\nu_\alpha = 10.2$ eV is the difference in energy between the $n = 2$ state and the ground state.

Recombination from the 2p state to the ground state happens via the emission of a Lyman-α photon. At the high gas densities present at $z \sim 1000$, however, most of these photons are rapidly reabsorbed by other hydrogen atoms, leading to no net change in the number of excited atoms. Consequently, recombination via Lyman-α emission can only take place as photons are lost from the Lyman-α resonance through cosmological redshifting. If we write the specific intensity in the Lyman-α resonance as I_{ν_α}, and assume that the radiation spectrum bluewards of the resonance is given by the Planck function, then we can write the rate at which photons are lost as

$$R = \frac{4\pi}{hc} H \left(I_{\nu_\alpha} - B_\nu \right) \text{ photons cm}^{-3} \text{ s}^{-1}. \tag{1.55}$$

Clearly, this is also the net recombination rate from 2p to 1s.

Equation 1.55 can be more conveniently written in terms of the occupation number $\mathcal{N}$, given by

$$\mathcal{N} = \frac{c^2}{2h\nu^3} I_\nu. \tag{1.56}$$

Using this, we find that

$$R = K^{-1} \left(\mathcal{N}_\alpha - \mathcal{N}_{pl} \right), \tag{1.57}$$

where

$$K = \frac{c^3}{8\pi\nu_\alpha^3} H^{-1} \tag{1.58}$$

and where $\mathcal{N}_\alpha$ and $\mathcal{N}_{pl}$ are the occupation numbers in the Lyman-alpha resonance and for a black-body spectrum respectively. The latter can be written as

$$\mathcal{N}_{pl} = \frac{1}{e^{h\nu_\alpha/kT} - 1} \tag{1.59}$$

which can be approximated as

$$\mathcal{N}_{pl} \simeq e^{-h\nu_\alpha/kT} \tag{1.60}$$

since $h\nu_\alpha \gg kT$. Hence

$$R = K^{-1} \left(\mathcal{N}_\alpha - e^{-h\nu_\alpha/kT} \right). \tag{1.61}$$

Finally, note that since Lyman-α transitions are rapid, we can write n_{2p}/n_{1s} as

$$\frac{n_{2p}}{n_{1s}} = \frac{3A_{21}\mathcal{N}_\alpha}{A_{21} + A_{21}\mathcal{N}_\alpha} \tag{1.62}$$

$$\simeq 3\mathcal{N}_\alpha, \tag{1.63}$$

where A_{21} is the spontaneous rate coefficient for the Lyman-α transition, and we have used the fact that $\mathcal{N}_\alpha \ll 1$. Moreover, since the excited levels are thermalized, we have

$$\frac{n_{2s}}{n_{1s}} = \frac{1}{3}\frac{n_{2p}}{n_{1s}} \tag{1.64}$$

and so

$$\frac{n_{2s}}{n_{1s}} = \mathcal{N}_\alpha. \tag{1.65}$$

We thus have three equations (nos. 1.50, 1.61 and 1.65) in terms of three unknowns (R, $\mathcal{N}_\alpha$ and n_{2s}) and it is a straightforward matter to solve for $\frac{\mathrm{d}n_\mathrm{H}}{\mathrm{d}t}$. We find that

$$\frac{\mathrm{d}n_\mathrm{H}}{\mathrm{d}t} = C \left(k_{2b} n_e^2 - \beta_e n_{1s} e^{-h\nu_\alpha/kT} \right) \tag{1.66}$$

where

$$C = \frac{1 + K\Lambda n_{1s}}{1 + K(\Lambda + \beta_e) n_{1s}}. \tag{1.67}$$

The term in brackets corresponds to the net recombination rate in the absence of the Lyman-α photons; these reduce the rate by a factor C.

At high redshift, $C \ll 1$ and recombination proceeds slowly. At lower redshift, $C \to 1$ and the rate coefficient increases. At the same time, however, the density is dropping sharply, and eventually the recombination timescale becomes longer than the expansion timescale. At this point the fractional ionization 'freezes out' at some small residual value. Several examples of this behaviour are shown in figure 1.1, for different cosmological models. A good fit to the residual ionization,

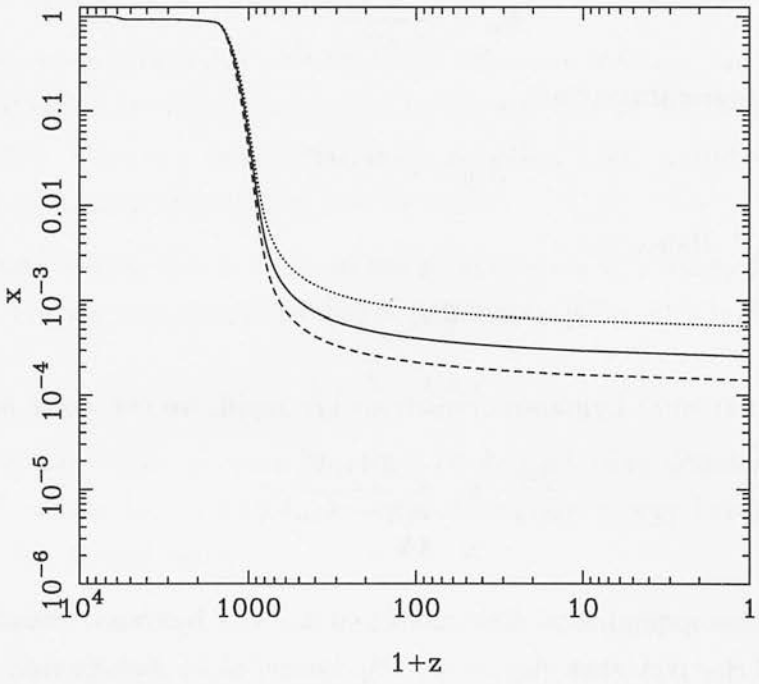

Figure 1.1: The fractional ionization as a function of redshift, for three cosmological models: $\Omega_b = 0.05$, $\Omega_m = 0.95$, $\Omega_v = 0$ (solid line); $\Omega_b = 0.05$, $\Omega_m = 0.25$, $\Omega_v = 0.7$ (dashed line); $\Omega_b = 0.025$, $\Omega_m = 0.975$, $\Omega_v = 0$ (dotted line). In all three models $h = 0.65$ and the helium mass fraction is 0.24. This plot, along with figure 1.2 was generated using data produced by RECFAST (Seager *et al.*, 1999).

x, obtained from the simplified treatment is (Peebles, 1993)

$$x = 1.2 \times 10^{-5} \frac{\Omega_m^{1/2}}{\Omega_b h}. \tag{1.68}$$

An accurate multi-level treatment finds a result only 10% smaller (Seager *et al.*, 2000).

Given $x(z)$, it is straightforward to calculate the evolution of the Thomson scattering optical depth

$$\tau = \int n(z)x(z)\sigma_T R(z)\,\mathrm{d}r \tag{1.69}$$

$$= \int n_0\sigma_T c\frac{x(z)(1+z)^2}{H(z)}\,\mathrm{d}z \tag{1.70}$$

where σ_T is the Thomson cross-section and n_0 is the baryon number density at the present day.

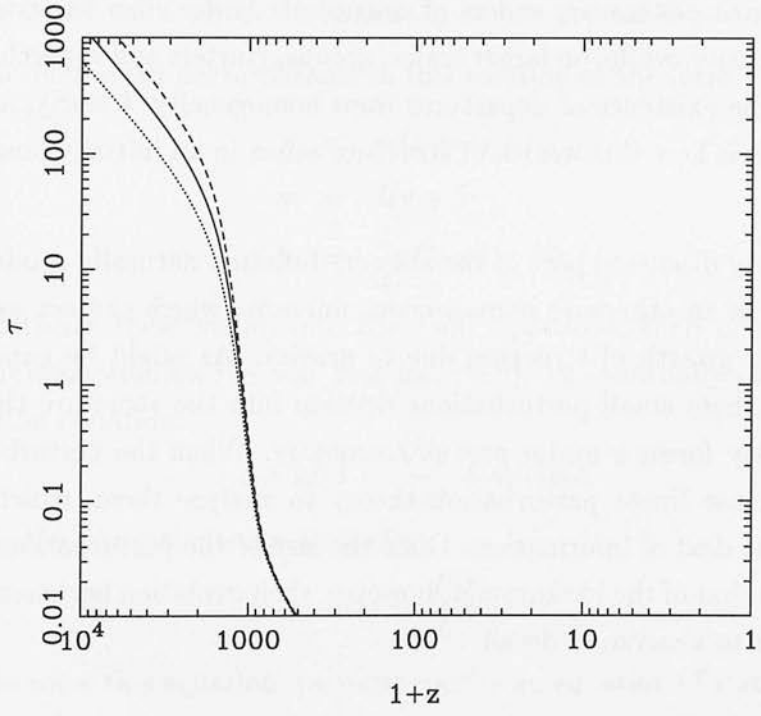

Figure 1.2: The Thomson scattering optical depth as a function of redshift, calculated for the same models as used in figure 1.1.

The evolution of τ for the same set of cosmological models is plotted in figure 1.2. We see that $\tau(z)$ is sensitive to the cosmological parameters only at high redshift, by which time it is already large. This is fortunate, as it implies that the redshift of last scattering – at which CMB photons are scattered for the last time – is independent of the cosmological parameters. Consequently, we can use the CMB as a probe of the physical conditions of the universe at a well-defined time in

its history. In particular, deviations from homogeneity at this time will create temperature anisotropies in the CMB, and measurements of the power spectrum of these anisotropies can be used to constrain the power spectrum of density fluctuations in the early universe. This is a large and ongoing area of research; for a good review of the basic physics see Hu *et al.* (1995).

1.2 The growth of structure

The standard model is a good description of the universe on scales on which we can regard it as effectively homogeneous. However, when we look at the universe on smaller scales, we find that it is largely *inhomogeneous*. Galaxies have mean densities that are orders of magnitude larger than the cosmological background density, while, on larger scales, groups, clusters and superclusters also demonstrate the existence of departures from homogeneity. Clearly, an obvious question to ask is how this wealth of structure arises in an initially homogeneous universe.

We have already discussed part of the answer. Inflation naturally produces small perturbations in an otherwise homogeneous universe, which can act as seeds for the subsequent growth of structure due to gravity. As might be expected, the study of how these small perturbations develop into the structure that we see around us today forms a major part of cosmology. When the perturbations are small, we can use linear perturbation theory to analyze them, which provides us with a great deal of information. Once the size of the perturbations becomes comparable to that of the background, however, their evolution becomes nonlinear and far harder to analyze in detail.

1.2.1 Linear theory

A wealth of literature exists on the study of cosmological perturbations in the linear regime, far more than I can do justice to in this brief review. [5] Rather than attempting to, I instead concentrate on the details of a particularly simple case – perturbations to a gas-filled, Einstein-de Sitter universe – which illustrates many of the basic concepts, and then go on to discuss the more general case in a far more qualitative fashion.

[5] Good general reviews are given by Efstathiou (1990) and Peebles (1980).

In an unperturbed universe, the fluid equations

$$\frac{\partial \rho}{\partial t} = -\nabla \cdot \rho \mathbf{v} \tag{1.71}$$

$$\frac{\partial \mathbf{v}}{\partial t} + (\mathbf{v} \cdot \nabla)\mathbf{v} = -\frac{\nabla p}{\rho} - \nabla \Phi \tag{1.72}$$

$$\nabla^2 \Phi = 4\pi G \rho, \tag{1.73}$$

have solutions

$$\rho_b = \rho_{b,0}(1+z)^3 \tag{1.74}$$

$$\mathbf{v} = H\mathbf{r} \tag{1.75}$$

$$\Phi_b = \frac{2}{3}\pi G \rho_b r^2, \tag{1.76}$$

where $\rho_{b,0}$ is the value of the mean density at the present day.

Let us consider small perturbations to this solution of the form

$$\rho = \rho_b(1+\delta) \tag{1.77}$$

$$\mathbf{v} = H\mathbf{r} + \delta\mathbf{v} \tag{1.78}$$

$$\Phi = \frac{2}{3}\pi G \rho_b r^2 + \delta\Phi \tag{1.79}$$

If we substitute these values into the fluid equations, then after a significant amount of manipulation (see e.g. Peebles, 1980), we eventually obtain the linear perturbation equation

$$\ddot{\delta} + 2H\dot{\delta} = \frac{\nabla^2 p}{\rho_b a^2} + 4\pi G \rho_b \delta, \tag{1.80}$$

where a is the dimensionless scale factor

$$a(t) = \frac{R(t)}{R(t_0)}. \tag{1.81}$$

In order to solve this equation, we must specify an equation of state. For an ideal gas, this can be written as

$$p = p_b + c_s^2 \rho_b \delta \tag{1.82}$$

where p_b is the unperturbed pressure and $c_s^2 = \partial p / \partial \rho$ is the sound speed. Substituting this into equation 1.80, we obtain

$$\ddot{\delta} + 2H\dot{\delta} = \left(\frac{c_s}{a}\right)^2 \nabla^2 \delta + 4\pi G \rho_b \delta, \tag{1.83}$$

which has solutions which are plane waves, $\delta = D(t)\, e^{i\mathbf{k}\cdot\mathbf{x}}$. Substituting this general solution into equation 1.83, we obtain

$$\ddot{D} + 2H\dot{D} = [4\pi G \rho_b - (c_s k/a)^2]D. \tag{1.84}$$

In the long wavelength limit (or if the pressure is negligible), this equation reduces to

$$\ddot{D} + 2H\dot{D} = 4\pi G\rho_b D, \tag{1.85}$$

which, in an Einstein-de Sitter universe, has power-law solutions

$$D_+ = t^{2/3}, \quad D_- = t^{-1}. \tag{1.86}$$

Since the initial perturbations are small, the decaying solution is of no interest and we do not consider it further. The growing solution, on the other hand, is of great importance as, given sufficient time to grow, it must eventually become non-linear.

In the small wavelength limit, pressure dominates self-gravity and equation 1.84 becomes

$$\ddot{D} + 2H\dot{D} = -(c_s k/a)^2 D. \tag{1.87}$$

If, additionally, expansion can be neglected, then the $\dot{D}$ term vanishes and the equation has solutions which are acoustic oscillations

$$D = D_0 \exp\left(\frac{-ic_s kt}{a}\right). \tag{1.88}$$

Including the effects of expansion in the adiabatic approximation alters this to

$$D = \frac{D_0}{(c_s a)^{1/2}} \exp\left(-i\int (c_s k/a)\,\mathrm{d}t\right) \tag{1.89}$$

as outlined in Peebles (1980).

The critical wavelength at which the character of the solution changes is given by

$$\lambda_J = c_s\left(\frac{\pi}{G\rho_b}\right)^{1/2}. \tag{1.90}$$

and is known as the *Jeans length* (Jeans, 1902). It is also useful to introduce the Jeans mass, defined as

$$M_J = \rho_b \lambda_J^3 \tag{1.91}$$

(although other definitions, differing by a small numerical factor, exist in the literature).

Roughly speaking, perturbations with mass greater than the Jeans mass will eventually collapse due to their own self-gravity (although pressure effects can delay this substantially for masses near M_J), while those with mass less than the Jeans mass are pressure-supported and do not collapse. In practice, this is an oversimplification, not least because M_J is not constant (see *e.g.* Gnedin & Hui, 1998). Nevertheless, it it a useful rule of thumb.

The general case

The preceding analysis is incomplete in a number of respects. In particular, it neglects the effects of radiation on the growth of perturbations. At the present epoch, $\rho_m \gg \rho_r$ and radiation has little effect, but as we go to higher redshift, the radiation density increases faster than the matter density; the two reach equality at

$$\begin{aligned} 1 + z_{eq} &= \frac{\rho_m}{\rho_r} \\ &= 24000 \Omega_m h^2. \end{aligned} \tag{1.92}$$

Moreover, prior to recombination, Thomson scattering couples the matter and radiation components tightly together.

This has two main effects. Firstly, the effective sound speed for a coupled mixture of matter and radiation is given by (Peacock, 1999)

$$c_s^2 = c^2 \left(3 + 2.25 \frac{\rho_m}{\rho_r} \right)^{-1}, \tag{1.93}$$

which for $\rho_r \sim \rho_m$ is a significant fraction of the speed of light. This means that the Jeans mass is very large

$$M_J \simeq 3.5 \times 10^{16} (\Omega_m h^2)^{-2} \, \mathrm{M}_\odot, \tag{1.94}$$

and the growth of perturbations on smaller scales is suppressed until matter and radiation decouple.

Secondly, perturbations on these smaller scales are actually erased, through a process known as Silk damping (Silk, 1968). Photons diffuse from overdense to underdense regions, and due to the tight coupling, the matter is carried along with them. The effect is to remove perturbations on scales smaller than the comoving damping length (Peacock, 1999)

$$\lambda_s = 16.3 (1 + z)^{-5/4} (\Omega_b^2 \Omega h^6)^{-1/4} \, \mathrm{Gpc} \tag{1.95}$$

This is comparable to the Jeans length at recombination.

Thus, in a purely baryonic universe, galaxies form through the fragmentation of larger structures; this is often known as 'top-down' structure formation. Unfortunately, the predictions of purely baryonic models conflict with observation in a

number of ways. In particular, it is possible to show that in order to produce the large-scale structure observed at the present day, density perturbations of amplitude $\delta \sim 10^{-3}$ are required at the time of last-scattering. These would produce temperature anisotropies $\Delta T/T \sim 3 \times 10^{-4}$, more than an order of magnitude larger than those actually observed (Silk & Wilson, 1981).

As mentioned previously, the solution to this problem is non-baryonic dark matter. As this does not interact electromagnetically, it has less effect on the radiation component, and dark matter perturbations of amplitude $\delta \sim 10^{-3}$ are consistent with the observed temperature anisotropies (Bond & Efstathiou, 1984). Moreover, small-scale perturbations in the dark matter are not suppressed by radiation pressure or erased by Silk damping. Rather, the minimum scale is set by dark matter 'free-streaming'. This occurs because weakly-interacting dark matter is effectively collisionless, and perturbations on scales smaller than $\lambda_{\rm fs} \sim \sigma t$, where σ is the velocity dispersion, are erased by the particles streaming away.

Various different models have been suggested for this non-baryonic dark matter (*e.g.* Bond *et al.*, 1980; Shafi & Stecker, 1984; Blumenthal *et al.*, 1984; Schaeffer & Silk, 1988), but for the past decade the leading model(s) have been based around cold dark matter (CDM) [6]. This has a thermal velocity which in cosmological terms is effectively zero, and the corresponding free-streaming scale is very small. Thus, to all intents and purposes, structure forms on all scales within the dark matter in CDM models. Prior to decoupling, small-scale structure is suppressed in the baryons, but after decoupling the baryons quickly 'catch up' with the dark matter, producing baryonic perturbations on all scales larger than the Jeans mass, which at this epoch is of order 10^6 $M_\odot$.

Current models thus consist of three components – baryons, CDM and radiation – with coupled perturbations. Accurate calculations should also take account of the effects of GR, rather than adopting a Newtonian approximation. Hence, although analytical calculations of the growth of perturbations are possible given certain simplifying approximations (Hu & Sugiyama, 1995), high-precision results require a numerical treatment. This generally involves solution of the covariant Boltzmann equation for all three components; an example of a state-of-the-art approach is given by the CMBFAST code of Seljak & Zaldarriaga (1996).

[6] Although there are some indications that CDM models may break down on small scales (Moore, 2001).

The cosmological density field

We have studied above how individual perturbations evolve, but in reality the density perturbation field

$$\delta(\mathbf{x}) \equiv \frac{\rho(\mathbf{x}) - \rho_b}{\rho_b} \tag{1.96}$$

will be made up of a superposition of perturbations on many different scales. Moreover, we generally do not have enough information to predict the precise distribution of these perturbations in space; indeed, for perturbations produced by inflation, we know that this distribution is random. Thus, any comparison between predicted and real density fields must be done in a statistical sense. For example, we can calculate the two-point correlation function $\xi(r)$ by averaging over an ensemble of realisations of the density field

$$\xi(\mathbf{r}) \equiv \langle \delta(\mathbf{x}) \delta(\mathbf{x} + \mathbf{r}) \rangle, \tag{1.97}$$

and using the statistical isotropy of the field to argue that $\xi(\mathbf{r}) \equiv \xi(r)$. This can then be compared with the observed two-point function, defined as the excess probability of finding two mass elements (eg galaxies) separated by a distance r

$$\mathrm{d}P = \rho^2 [1 + \xi(r)] \, \mathrm{d}V_1 \, \mathrm{d}V_2. \tag{1.98}$$

If our model for the genesis and growth of the density perturbations is correct, then the two should be the same [7]. Similar comparisons can be performed for higher-order correlation functions.

We saw in our simple treatment of perturbation growth that the basic solution had the form of a plane wave. This result remains true in the general case, and leads to the important conclusion that we can decompose an arbitrary perturbation into a set of plane-wave modes, which will then evolve independently while they remain in the linear regime. This suggests that instead of working in terms of $\xi(r)$, whose evolution is complicated, we instead work in terms of its Fourier transform, $P(k)$, known as the power spectrum. This is defined by (Peacock, 1999)

$$\xi(r) = \frac{V}{(2\pi)^3} \int P(k) \frac{\sin kr}{kr} 4\pi k^2 \, \mathrm{d}k. \tag{1.99}$$

To avoid the mathematical difficulties involved with performing Fourier transforms of infinite fields, we have assumed that the perturbations are periodic

[7] This assumes that the density field is ergodic, *i.e.* that we can convert from ensemble averages to volume averages and vice versa (Adler, 1981).

within a box of volume V; the result appropriate to an infinite universe can then be obtained by taking the limit as $V \to \infty$. [8].

To fully describe an arbitrary density field, we would require not only the two-point function (or power spectrum), but also the rest of the infinite set of correlation functions. However, if the density field is a Gaussian random field (Bardeen *et al.*, 1986), then all of these either vanish or can be specified in terms of the two-point function – knowledge of $P(k)$ is sufficient to completely determine the statistical properties of the field. We might expect on quite general grounds that the initial density field should be Gaussian – the central limit theorem implies that the sum of a large number of random variables approaches a Gaussian distribution – and inflationary models universally predict that it should be so.

Moreover, the Gaussian nature of the field is unaffected by its evolution as long as all of its modes remain in the linear regime. Non-linear evolution, on the other hand, inevitably introduces non-Gaussianity into the density field, as δ has a minimum bound, but no maximum bound. However, if clustering grows in a 'bottom-up' fashion, as in CDM, then even strongly non-linear clustering has little effect on the behaviour of longer wavelength modes (Peebles, 1974). Thus, while the true density field may no longer be Gaussian, if we smooth it on a scale greater than that of the non-linear clustering, the resulting smoothed field *will* be Gaussian. Consequently, $P(k)$ remains a useful tool for describing large-scale structure right up until the present epoch, as long as we recall that it is directly applicable only on the largest scales.

In order to determine $P(k)$ we start with the inflationary prediction. This is generally of power law form

$$P(k) = Ak^n \tag{1.100}$$

with $n \sim 1$ in plausible models; the particular case $n = 1$ is known as the scale-invariant or Harrison-Zeldovich spectrum (Harrison, 1970; Zeldovich, 1972) The normalisation, A, is not strongly constrained by inflation and is generally determined either from the amplitudes of the CMB anisotropies (Liddle & Lyth, 1993) or from the rich cluster abundance (Henry & Arnaud, 1991).

Subsequently, individual modes evolve as outlined in the previous section; a good discussion of the effect this has on the power spectrum is given in Padmanabhan

[8]Note that since $P(k)$ is dimensionless, the factor of V^{-1} that we pick up when performing the integration will cancel with the factor of V in the normalization; $\xi(r)$ is independent of V.

(1993). These evolutionary effects are normally summarized in a single quantity, the transfer function $T(k)$. This is defined as

$$T(k) \equiv \frac{\delta_k(z=0)}{\delta_k(z)D(z)} \tag{1.101}$$

where $D(z)$ is the linear growth factor, implying that

$$P(k;z=0) = T(k)^2 D(z)^2 P(k;z). \tag{1.102}$$

Taking z to be sufficiently large that $P(k;z)$ still has its primordial value, and including the factor D^2 in the (undetermined) normalisation, this can be written as

$$P(k) \propto T(k)^2 k^n. \tag{1.103}$$

A useful fitting formula for the transfer function in CDM models is given by (Bardeen *et al.*, 1986)

$$T(k) = \frac{\ln(1+2.34q)}{2.34q}[1+3.89q+(16.1q)^2+(5.46q)^3+(6.71q)^4]^{-1/4} \tag{1.104}$$

where

$$q \equiv \frac{k/h\ \mathrm{Mpc}^{-1}}{\Gamma} \tag{1.105}$$

and the shape parameter Γ is given by (Sugiyama, 1995)

$$\Gamma = \Omega_m h \exp\left[-\Omega_b(1+\sqrt{2h}/\Omega_m)\right]. \tag{1.106}$$

This fitting formula is adequate as long as the baryon fraction is small; if this is not the case, then the high-k end of the power spectrum can develop significant oscillations (Meiksin *et al.*, 1999).

Armed with $P(k)$ we can now construct any of the statistical properties of the field. One of the most useful, which we will encounter again in section 1.2.4, is the variance of the smoothed field, $\sigma(R)^2$. This is given by (Peacock, 1999)

$$\sigma(R)^2 = \frac{V}{(2\pi)^3}\int P(k)|W(kR)|^2 \mathrm{d}^3\mathrm{k}, \tag{1.107}$$

where $W(kR)$ is a filter function with characteristic scale R. Common choices for the filter function are the top-hat filter

$$W_{\mathrm{TH}} = \frac{3}{(kR)^3}(\sin kR - kR\cos kR), \tag{1.108}$$

the Gaussian filter

$$W_{\rm G} = e^{-(kR)^2/2} \tag{1.109}$$

or the sharp k-space filter

$$W_{\rm k} = \begin{cases} 1 & k < k_{\rm F} \\ 0 & k > k_{\rm F} \end{cases} \tag{1.110}$$

where $k_{\rm F} = (9\pi/2)^{1/3} R^{-1}$ (Lacey & Cole, 1993).

A common application of equation 1.107 is the determination of the normalisation of $P(k)$. This can be done by calculating σ_8, the rms fluctuation of a density field smoothed with a top hat filter of radius $8h^{-1}$Mpc, and adjusting the normalisation until it matches the observed value, which is of order unity.

Unless otherwise stated, the cosmological model adopted for the remainder of this thesis is the ΛCDM concordance model of Wang *et al.* (2000). In this model, the cosmological parameters are

$$\begin{array}{ll} \Omega_\Lambda = 0.67 & \Omega_m = 0.33 \\ \Omega_b = 0.041 & h = 0.65 \\ n = 1.0 & \sigma_8 = 0.9 \end{array}$$

This is an updated version of the model presented in Ostriker & Steinhardt (1995), and is consistent with the current observational bounds.

1.2.2 Non-linear evolution

Eventually, perturbations will grow to a point at which linear theory breaks down. This can be delayed somewhat through the use of higher-order perturbation theory (Bouchet, 1996), but when $\delta \sim 1$ perturbation theory of any order must break down. To make predictions about the non-linear epoch requires the use of other techniques.

One of the most useful such techniques is the spherical collapse model. As the name suggests, this deals with the evolution of spherically symmetric perturbations. Birkhoff's theorem (Birkhoff, 1923) implies that the dynamics of a spherical perturbation at any radius are entirely determined by the mean density within that radius. In this case, evolution equations are easy to derive; the proper radius depends on time through the parametric equations

$$r = A(1 - \cos\theta) \tag{1.111}$$

$$t = B(\theta - \sin\theta), \tag{1.112}$$

where $A^3 = GMB^2$, and M is the mass enclosed.

We immediately see that there are two interesting epochs in the growth of the perturbation. The first occurs when $\theta = \pi$, and r reaches its maximum value. We refer to this point as *turnaround* – in proper coordinates, it is the point at which collapse truly begins. The other occurs at $\theta = 2\pi$, when the equations predict that the sphere will have collapsed to a singularity. This prediction is a consequence of our assumption of perfect spherical symmetry, and will not occur for any real perturbation. Rather, aspherical motions will eventually halt the collapse, even if pressure forces remain negligible. The final state of the system can then be determined from the virial theorem. If we assume that energy is conserved during the collapse, then the total energy of the virialized perturbation, given by the sum of its potential energy V and kinetic energy K, must be equal to its potential energy at turnaround [9]

$$V + K = V_{\rm ta}. \tag{1.113}$$

However, the virial theorem tells us that

$$V = -2K, \tag{1.114}$$

so the potential energy at virialisation is related to that at turnaround by

$$V = 2V_{\rm ta}. \tag{1.115}$$

Now, for a uniform spherical perturbation, the potential energy is given by

$$V = \frac{3}{5}\frac{GM^2}{R}, \tag{1.116}$$

so equation 1.115 implies that

$$R_{\rm vir} = \frac{1}{2}R_{\rm ta}, \tag{1.117}$$

i.e. the virialized perturbation has half the radius (and hence eight times the density) of the perturbation at turnaround. Note that this analysis assumes that the effect of the cosmological constant (if present) is negligible; Lahav *et al.* (1991) consider the case when it is not.

This toy model allows us to calculate the density contrast of the virialized perturbation. If the background universe is well-described by the Einstein-de Sitter

[9] Note that we also assume that the perturbation is small enough that the use of Newtonian gravity in place of GR is justified.

model, we find that $\delta_{\rm vir} = 18\pi^2 \simeq 178$; this is to be compared with a linearly-extrapolated density contrast of only $\delta_{\rm lin} \simeq 1.69$. Analogous results can be obtained for more general background universes; if $\Omega = 1$ then (Bryan & Norman, 1998)

$$\delta_{\rm vir} = 18\pi^2 + 82x - 39x^2, \tag{1.118}$$

whereas if $\Omega < 1$ and $\Lambda = 0$, then

$$\delta_{\rm vir} = 18\pi^2 + 60x - 32x^2. \tag{1.119}$$

In both cases, $x \equiv \Omega_m(z) - 1$. The density contrast derived in this way is in reasonable agreement with that determined from large N-body simulations (Cole & Lacey, 1996).

It is quite straightforward to generalize the spherical collapse model to the case of a homogeneous ellipsoid (see section 20 of Peebles, 1980), but in this case Birkhoff's theorem no longer applies and the internal evolution of the ellipsoid will inevitably be affected by the large-scale tidal field. Consequently, the model has not found a great number of applications (although see Sheth *et al.*, 2001 or Monaco, 1997).

Other analytical methods also exist which can be used for study of the non-linear epoch – for example, the Zeldovich approximation Zeldovich (1970) and its extensions (Gurbatov *et al.*, 1989; Matarrese *et al.*, 1992; Bagla & Padmanabhan, 1994) – but these are generally more suited for the study of planar and filamentary structures, such as the Lyman-α forest, than for investigating the quasi-spherical structures involved in galaxy formation. To gain more insight into these structures generally requires the use of numerical simulations.

1.2.3 Numerical simulation

There are two main types of numerical simulation in widespread use in the study of galaxy formation: N-body simulations and hydrodynamical simulations.

N-body simulations do not include a gas component, concentrating instead on the dark matter. The evolution of the dark matter is determined purely by gravity and the baryons have little effect so long as $\rho_{\rm dm} \gg \rho_g$. A number of different strategies exist for numerical solution of the equations of motion, and a good overview of the field is given in Bertschinger (1998). Clearly, neglect of

the baryons limits the type of situations which can be studied, but it also allows far greater resolution and/or dynamical range for a given amount of computing power than would be possible for a hydrodynamical code.

N-body simulations have provided us with a number of important insights into galaxy formation. In particular, they are an excellent tool for studying the clustering properties of galaxies in CDM structure formation models (Colberg *et al.*, 2000), as long as some assumption is made as to how light traces mass. They have also proved successful in the study of the structure of individual dark matter halos, although uncertainties remain in the structure of the baryon-dominated central regions (Navarro *et al.*, 1997; Moore *et al.*, 1999).

Hydrodynamical simulations, which trace the evolution of the baryons as well as the dark matter, should in theory give us a much better picture of galaxy formation. In practice, this is less true than one might have hoped. This is primarily due to our limited understanding of star formation and stellar feedback, and how best to include their effects into the simulations (see Thacker & Couchman, 2000, for a recent view of the subject); additionally, the large dynamical range that is required in hydrodynamical simulations leads to a number of computational difficulties, although the development over the past few years of sophisticated adaptive-mesh codes (Pearce & Couchman, 1997; Truelove *et al.*, 1998) seems likely to alleviate this problem.

Nevertheless, as we shall see in later sections, hydrodynamical simulations play an important role in developing our insight into the details of galaxy formation, even if their specific predictions often have to be taken with a pinch of salt.

1.2.4 The Press-Schechter formalism

Whilst we could run an N-body simulation every time we needed to know the halo mass function, this would be remarkably inefficient. Far better would be to have some analytic or semi-analytic method by which to calculate it.

The Press-Schechter formalism is one such method, probably the one in widest use. It was originally suggested by Press & Schechter (1974), reformulated and clarified by Bond *et al.* (1991) and subsequently extended by Lacey & Cole (1993). It allows one to obtain the mass function of collapsed objects (such as galaxies) from the statistics of the *linear* density field.

Let us define the mass function $n(M,z)\,\mathrm{d}M$ to be the comoving number density of collapsed objects with masses in the range M to $M+\mathrm{d}M$ at a redshift of z. If the total mass fraction in objects of mass greater than or equal to M is $f(M,z)$ then

$$n(M,z) = \frac{\rho_b}{M}\left|\frac{df}{dM}\right|. \tag{1.120}$$

Thus, to determine $n(M,z)$ we must determine $f(M,z)$. In their seminal paper, Press & Schechter suggested a way that this could be done based only on the statistics of the linearly extrapolated density field.

We begin by ignoring the details of the non-linear evolution of the density field; rather, we suppose that the density contrast at any given point grows linearly until it reaches some critical value δ_c, at which point it instantaneously becomes part of some collapsed object. Consideration of the spherical collapse model suggests that a suitable critical density is $\delta_c = 1.69$, the linearly extrapolated overdensity of a spherical perturbation at the moment of collapse. The assumption of linear evolution implies that the form of the density field is preserved, so that if is initially a Gaussian random field (Bardeen *et al.*, 1986), it will remain so. This significantly simplifies the resulting analysis.

We smooth the field with a filter function of characteristic scale length R, which we can associate with a mass $M(R)$; clearly $M \propto R^3$, with the constant of proportionality depending upon the choice of filter function. We then apply the density threshold criterion to this smoothed field. Points with $\delta = \delta_c$ are considered to have just collapsed and are associated with objects of mass M. Points with $\delta > \delta_c$ will have $\delta = \delta_c$ when smoothed on some larger scale, and thus will correspond to a more massive object. By repeating this procedure on a range of different scales, we can obtain the collapsed fraction corresponding to each scale and hence obtain $f(M,z)$. If we do this for a Gaussian density field, we find that

$$f(M,z) = \frac{1}{2}\,\mathrm{erfc}\left[\frac{\delta_c}{\sqrt{2}\sigma(\mathrm{M})}\right] \tag{1.121}$$

where σ^2 is the variance of the smoothed field.

This, in essence, is the formalism as originally proposed by Press & Schechter. Unfortunately, in suffers from a normalisation problem. As we can see if we let $M \to 0$ in equation 1.121, it accounts for only half of the mass in the Universe. The problem arises because of the way that underdense regions are treated. The procedure does not allow for the possibility of a large overdense region containing

smaller underdense regions, which, although unable to collapse on their own, would be incorporated in the collapse of the larger region. This is known as the cloud-in-cloud problem. Press & Schechter 'solved' the problem by multiplying by a factor of two, with little real justification.

A better resolution to this problem comes from the work of Bond *et al.* (1991), who reformulated the method in terms of excursion sets. The excursion set approach begins by smoothing on the largest possible scale, where homogeneity implies that δ is zero. If the smoothing scale is then reduced, the density contrast at a given point begins to move away from zero, with a trajectory which depends upon the spatial location and the properties of the density field. If the point chosen forms part of a collapsed object then eventually a smoothing scale is reached for which $\delta = \delta_c$, and we can identify the point as part of an object with a mass corresponding to the smoothing scale. Moreover, since we are working *downward* from the largest possible mass, we know for certain that the point is not contained within some larger object. In this way we hope to avoid the cloud-in-cloud problem.

In practice, of course, only the statistical properties of the density field are known. Thus, rather than calculate the actual trajectories δ throughout space, we calculate the trajectories at a representative point for an ensemble of realizations of the density field, and subsequently obtain their statistical properties by averaging over the ensemble.

The detailed properties of the trajectories depend upon the form of filter function chosen. If we use a sharp k-space filter, then the increments and decrements to $\delta(M)$ as we decrease M are simply the amplitudes of the various Fourier modes that are added in. Since these are uncorrelated, the resulting trajectory is a Brownian random walk. For other types of filter this is not the case, and the subsequent mathematics is rather more complicated – see Peacock & Heavens (1990) or Bond *et al.* (1991) for more details.

To simplify the resulting mathematics, we remove the time dependence from the density field by making the density threshold time dependent:

$$\delta_c(t) = \frac{\delta_{c,0}}{D(t)} \tag{1.122}$$

where $D(t)$ is the linear growth factor and $\delta_{c,0} = 1.69$.

The number density of trajectories with overdensities in a particular interval

$\delta, \delta + \mathrm{d}\delta$ depends upon the smoothing scale, or alternatively upon the variance σ^2, since σ is a monotonic function of M. If we denote this number density by $F(\delta, \sigma)$, then

$$\frac{\partial F}{\partial \sigma^2} = \frac{1}{2} \frac{\partial^2 F}{\partial \delta^2}. \tag{1.123}$$

Thus the trajectories are governed by a simple diffusion equation. We can incorporate the density threshold by placing a barrier at $\delta = \delta_c$ that absorbs any trajectories that attempt to cross it. In this case the diffusion equation has the unique solution (Chandrasekhar, 1943)

$$F(\delta, \sigma, \delta_c)\ \mathrm{d}\delta = \frac{1}{\sqrt{2\pi\sigma^2}} \left[\exp\left(-\frac{\delta^2}{2\sigma^2} \right) - \exp\left(-\frac{(\delta - 2\delta_c)^2}{2\sigma^2} \right) \right] \mathrm{d}\delta. \tag{1.124}$$

The probability that a particular trajectory is absorbed in the interval $\sigma, \sigma + \mathrm{d}\sigma$ must equal the reduction in the number of trajectories that survive below the barrier. Hence the probability $p(\sigma, \delta_c)\,\mathrm{d}\sigma$ that a trajectory is absorbed in this interval is simply given by

$$p(\sigma, \delta_c)\,\mathrm{d}\sigma = \sqrt{\frac{2}{\pi}} \frac{\delta_c}{\sigma^2} \exp\left[-\frac{\delta_c^2}{2\sigma^2} \right] \mathrm{d}\sigma. \tag{1.125}$$

Since trajectories that are absorbed correspond to objects that have collapsed, it is easy to see that $f(M, z)$ is given by

$$f(M, z) = \int_0^{\sigma} p(\sigma, \delta_c)\,\mathrm{d}\sigma = \mathrm{erfc}\left[\frac{\delta_c}{\sqrt{2}\sigma} \right]. \tag{1.126}$$

Thus, we obtain the Press-Schechter result, but this time with the correct normalization. The resulting mass function is given by

$$n(M)\,\mathrm{d}M = \sqrt{\frac{2}{\pi}} \frac{\rho_b}{M^2} \frac{\delta_c}{\sigma} \left| \frac{\mathrm{d}\ln\sigma}{\mathrm{d}\ln M} \right| \exp\left(-\frac{\delta_c^2}{2\sigma^2} \right) \mathrm{d}M. \tag{1.127}$$

Unfortunately, this result is only obtained for the particular choice of a sharp k-space filter; other choices of filter lead to a different normalisation and to systematic differences in the low-mass slope. Also, N-body simulations show that the correlation between predicted and actual masses computed on a particle-by-particle basis is poor (Bond *et al.*, 1991). Nevertheless, the mass function predicted by the Press-Schechter formalism actually proves to be a good fit to that measured from N-body simulations (Lacey & Cole, 1994; Jang-Condell & Hernquist, 2001), and it is a widely used approximation. If greater accuracy is

required then a number of improvements to the Press-Schechter approach exist – see, for example van Kampen *et al.* (1999). Many of these improvements, along with a number of alternative approaches to the determination of a mass function, are discussed at length in the review of Monaco (1998).

1.3 Galaxy formation

Although the various techniques discussed in the previous section tell us a great deal, they do not address the essential aspect of galaxy formation – the formation of stars. Despite decades of work, and a huge amount of observational data, many details of this process remain to be understood even for stars forming at the present day and it is clear that any theory of primordial star formation will be incomplete. Nevertheless, it is possible to identify some necessary conditions for star formation. One of the most important of these is the requirement that gas be self-gravitating. For this to be the case, the gas density must exceed the dark matter density, and this can only happen if the gas is able to dissipate energy and collapse. Understanding how this occurs gives us important insights into the efficiency and nature of primordial star formation.

The free-streaming length of CDM is very small, allowing structure to form on all scales. Moreover, structure forms hierarchically, with smaller objects forming (on average) before larger ones. The process is not entirely self-similar, but it is still true that clustering at late times broadly resembles that at early times with a change of mass scale.

The same is not true for the gas. Pressure forces prevent collapse on scales less than the Jeans mass, and the gas component remains smoothly distributed until a relatively low redshift, with the first collapsed structures – the first protogalaxies – forming only once sufficiently deep potential wells exist in the dark matter.

Now, if the dark matter is distributed as a Gaussian random field, then such potential wells, corresponding to high peaks of the density field, will always exist *somewhere* in the universe. In practice, however, very rare peaks are of no interest, as the probability of finding one in our Hubble volume will be extremely small. A possible way of defining the epoch of galaxy formation is by determining when such peaks become relatively common. For example, figure 1.3 shows the evolution with redshift of the mass scale $M_{\rm NL}$ of a 3σ peak of the density

field. [10] Use of the Press-Schechter formalism allows us to determine the fraction of the total mass that is in objects of mass $M_{\rm NL}$ or greater; for a 3σ peak this is approximately 0.27%. Also plotted is the evolution with redshift of $M_{\rm J}$. The two intersect at $z \simeq 25$, at which time $M_{\rm NL} \simeq 2.2 \times 10^4\ {\rm M}_\odot$.

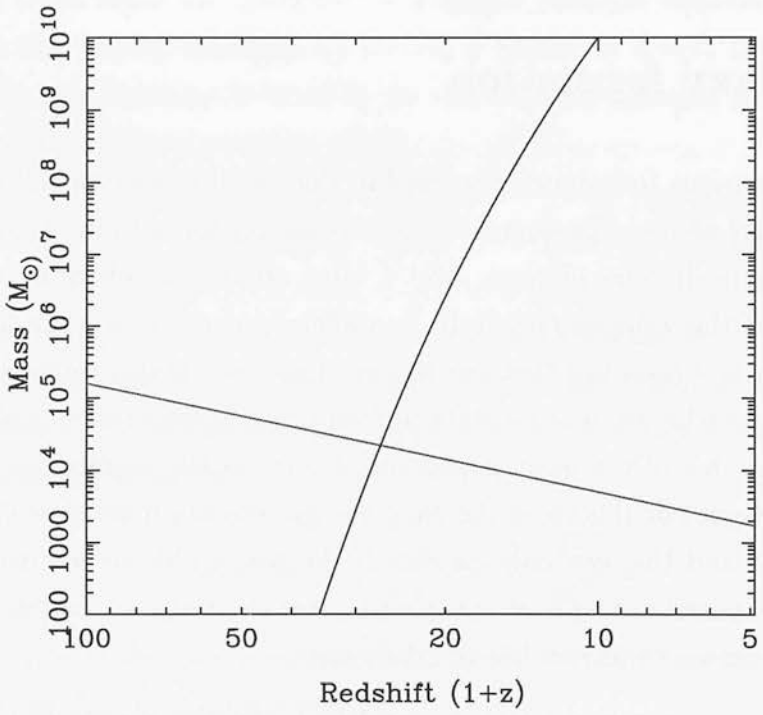

Figure 1.3: The evolution with redshift of the Jeans mass (red curve) and the mass-scale of a 3σ peak (blue curve) in our fiducial ΛCDM cosmological model.

Once deep enough potential wells exist that the gas can begin to collapse, what then? If the collapsing gas can from the outset dissipate heat effectively, then collapse may continue for a long time. If, on the other hand, dissipation is initially small, then it is reasonable to suppose that the infalling gas will be shocked and/or adiabatically heated to a temperature $T \sim T_{\rm vir}$, the virial temperature of the halo. This can be obtained by assuming that the kinetic energy of the virialized halo is entirely in the form of internal motions; in other words that

$$K = U_{\rm th} \equiv \frac{3}{2}\frac{kT_{\rm vir}}{\mu m_{\rm H}} M, \tag{1.128}$$

where M is the mass of the protogalaxy, and μ is the mean molecular mass. Now,

[10] Altering the cosmological model changes the quantitative evolution of $M_{\rm NL}$, but the qualitative evolution remains the same.

the virial theorem implies that

$$K = -\frac{V}{2} \tag{1.129}$$

so we can write the virial temperature in terms of the potential energy of the protogalaxy, V, as

$$T_{\rm vir} = \frac{\mu m_{\rm H}}{3k}\left(\frac{-V}{M}\right). \tag{1.130}$$

The value of V depends upon the density profile of the protogalaxy. For a uniform density profile we find that

$$T_{\rm vir} = \frac{2}{5}\frac{GM\mu m_{\rm H}}{kR_{\rm ta}} \tag{1.131}$$

while a singular isothermal sphere gives us

$$T_{\rm vir} = \frac{6}{5}\frac{GM\mu m_{\rm H}}{kR_{\rm ta}}. \tag{1.132}$$

Finally, approximating the density profile as a truncated isothermal sphere gives us (Shapiro *et al.*, 1999)

$$T_{\rm vir} = \frac{2\alpha}{5(\alpha-2)}\frac{GM\mu m_{\rm H}}{kR_{\rm ta}} \tag{1.133}$$

where $\alpha = 3.73$. Consideration of hydrodynamical simulations such as those performed by Abel *et al.* (2000) shows that either of the isothermal sphere profiles are reasonable approximations. However, the latter is more self-consistent, and avoids the central singularity of the former; unless otherwise noted, I adopt it hereafter.

Cooling

The subsequent behaviour of the gas will be determined by its ability to cool. As first demonstrated by Rees & Ostriker (1977), there are essentially two fates for the gas. If it is unable to cool rapidly, then it will remain in the form of a large, pressure supported cloud, evolving quasi-statically unless significantly perturbed. On the other hand, if cooling is rapid then the period of pressure support will be short or absent and the gas will continue to collapse. As it does so, the value of the Jeans mass within the cloud will decrease, and gravitational fragmentation (leading, one generally assumes, to the eventual formation of stars) becomes possible (Palla *et al.*, 1983).

We can determine which of these fates awaits the gas in a particular protogalaxy by comparing its cooling timescale

$$t_{\rm cool} = \frac{3}{2}\frac{nkT}{\Lambda(T)n^2}, \tag{1.134}$$

(where $\Lambda(T)$ is the cooling function) with its gravitational collapse timescale, or 'free-fall' time

$$t_{\rm ff} = \left(\frac{3\pi}{32G\rho}\right)^{1/2}. \tag{1.135}$$

If $t_{\rm ff} \gg t_{\rm cool}$ then the gas cools rapidly, sinks to the centre of the dark matter halo and begins to form stars. On the other hand, if $t_{\rm cool} \gg t_{\rm ff}$ then quasi-static evolution quickly sets in, and the gas remains in an extended configuration.

How then does primordial gas cool? At high temperatures, cooling by atomic line excitation is efficient and the gas cools rapidly. Below $T \sim 10^4$ K, however, atomic cooling becomes negligible and an alternative coolant is required. This temperature scale corresponds to a virial mass $M \simeq 5 \times 10^8 h^{-1}(1+z)^{-3/2}\,{\rm M}_\odot$, significantly larger than the Jeans mass, so an understanding of the thermal behaviour of primordial gas at lower temperatures will clearly be important. The chemical and thermal evolution of low temperature primordial gas have been discussed by a number of authors (Black, 1991; Dalgarno & Lepp, 1987; Galli & Palla, 1998; Abel *et al.*, 1997; Stancil *et al.*, 1998), and are also discussed at length in chapter 2 of this thesis. The dominant coolant low temperature coolant proves to be molecular hydrogen, H_2.

In order to understand the dynamical evolution of a protogalaxy, we must first understand its chemical evolution. In particular, we need to know whether enough H_2 can form to cool the gas effectively. If we were to look at primordial protogalaxies as being analogous to present day molecular clouds – they do, after all, have similar masses and sizes – then we would predict that a large quantity of H_2 would form in the gas. However, this would be to neglect a fundamental difference between the chemistry of primordial and metal-enriched gas. In molecular clouds, H_2 forms predominantly on the surface of dust grains, at a rate (Draine & Bertoldi, 1996)

$$R = 6 \times 10^{-18} T^{1/2} n_{\rm H}^2\ {\rm cm}^{-3}\ {\rm s}^{-1} \tag{1.136}$$

for dust of similar composition to that in the Milky Way. [11] This is an efficient

[11] The direct association of two hydrogen atoms to form H_2 is strongly forbidden, and proceeds at a negligible rate.

process, and can rapidly convert the bulk of the gas to molecular form. In primordial gas, however, there is no dust. In this case, the main routes of H_2 formation are two sets of gas phase reactions:

$$H + e^- \rightarrow H^- + \gamma \quad (1.137)$$

$$H^- + H \rightarrow H_2 + e^- \quad (1.138)$$

and

$$H + H^+ \rightarrow H_2^+ + \gamma \quad (1.139)$$

$$H_2^+ + H \rightarrow H_2 + H^+. \quad (1.140)$$

Formation via H^- generally dominates, being one to two orders of magnitude faster than formation via H_2^+. In both cases the limiting step is the formation of the molecular ion. The rate of this process is generally much smaller than that of formation on dust grains – the rate coefficients are comparable, but formation via the gas phase depends upon the electron (or proton) density, while formation via grains does not. Consequently, the gas phase reactions are typically several orders of magnitude slower; moreover, in the absence of a continuing source of ionization, the amount of H_2 that can form will be limited by recombination (Nishi & Susa, 1999).

Given these difficulties, it is reasonable to ask whether enough H_2 will *ever* form to cool the gas effectively. The amount that is required is simple to estimate. If we assume that all of the cooling is due to H_2, then setting $t_{\rm cool} = t_{\rm ff}$ implies that

$$x_{H_2} = 1.5 \times 10^{-31} \frac{T_{\rm vir}}{n^{1/2} \Lambda(T_{\rm vir})}, \quad (1.141)$$

where x_{H_2} is the required H_2 abundance and $\Lambda(T_{\rm vir})$ is the H_2 cooling function. In figure 1.4, the required H_2 abundance is plotted as a function of density and temperature. To put these values into context, recall that a collapsed object with an overdensity $\delta \sim 200$ will have a baryon number density $n \sim 1\,{\rm cm}^{-3}$ at redshift $z = 25$. The required H_2 abundance corresponding to this density has a strong temperature dependence, but for $T_{\rm vir} \sim 10^3$ K is of the order of a few times 10^{-4}, implying that a relatively small fractional abundance of H_2 will suffice to cool the gas. Nevertheless, this abundance is still orders of magnitude higher than that present in the IGM prior to collapse, which Stancil *et al.* (1998) have calculated to be 2.4×10^{-6}. If these first protogalaxies are to be able to cool, they must form the bulk of their H_2 during or after the initial collapse.

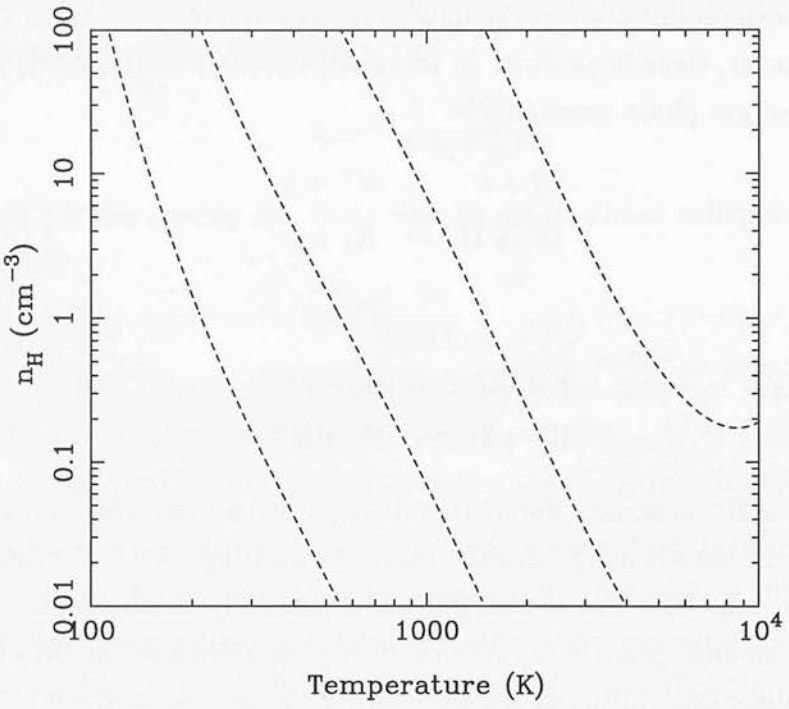

Figure 1.4: A contour plot of the H_2 abundance required for effective cooling; from top to bottom, contours correspond to fractional abundances of 10^{-5}, 10^{-4}, 10^{-3} and 10^{-2} respectively. Effective cooling is defined by the requirement that $t_{\rm cool} = t_{\rm ff}$. A uniform sphere model, with $\rho_{\rm vir} = 18\pi^2\rho_b$, is assumed for the density profile. The H_2 cooling function is from Galli & Palla (1998).

Various methods have been used to study the chemical evolution of forming protogalaxies. The most detailed, and hence the most informative, are 3D hydrodynamical simulations. However, these are computationally expensive, limiting both the amount of physics that can be included, and the range of initial conditions that can be studied. There is thus a strong motivation for supplementing them with simpler models.

The Tegmark model

One of the simplest models in widespread use is that of Tegmark *et al.* (1997). This makes use of an extremely simplified hydrodynamical model. The protogalaxy is approximated as a uniform sphere, whose density evolves as

$$\rho = \rho_b \exp\left(\frac{1.9A}{1 - 0.75A^2}\right) \tag{1.142}$$

(where $A = 1+z_{\rm vir}/1+z$) until it reaches some limiting density $\rho_{\rm lim} = \min(\rho_{\rm vir}, \rho_p)$. The latter is given by

$$\rho_p = \rho_b \left(\frac{T_{\rm vir}}{T_{\rm IGM}} \right)^{3/2} \tag{1.143}$$

and is an attempt to account for the effects of pressure support prior to virialisation. Equation 1.142 is an approximation to the density evolution within the spherical collapse model; it was first introduced by Tegmark (1994) and is accurate to within 5%. [12]

Gas in the protogalaxy is heated during collapse by adiabatic compression, and by shocks if $T < T_{\rm vir}$ at $z_{\rm vir}$. Meanwhile, cooling can proceed via the rotational lines of H_2, via the Compton scattering of CMB photons, or at high temperatures by Lyman-α cooling; note that this is substantially simplified compared to the model presented in chapter 2. The chemical modelling of H_2 is similarly limited, consisting of hydrogen recombination, the formation of H_2 via H^- or H_2^+, plus the destruction of these ions by the CMB.

Collapse is determined by a cooling criterion: only gas whose temperature satisfies

$$T(\eta z) < \eta T(z), \tag{1.144}$$

where $\eta = 0.75$, is considered to have cooled and collapsed. This is actually a less restrictive condition than the requirement that $t_{\rm cool} < t_{\rm ff}$ discussed above; for gas cooling at a roughly constant rate, it corresponds to requiring that $t_{\rm cool} \lesssim 10 t_{\rm ff}$. Given the uncertainties in the density modelling, however, it should still provide us with reasonable results.

The resulting model requires little computational power, and the T-z parameter space can be explored quickly to identify regions that can and cannot cool. Tegmark *et al.* find that at each redshift there exists a critical temperature $T_{\rm crit}$ such that protogalaxies with $T_{\rm vir} > T_{\rm crit}$ can cool efficiently, while those with $T_{\rm vir} < T_{\rm crit}$ cannot. The values of $T_{\rm crit}$ that they obtain are plotted in figure 1.5. The change in behaviour at $z \sim 50$ corresponds to the epoch at which pressure effects become important; at lower redshifts, $\rho_{\rm vir} \ll \rho_p$ and pressure does not limit the post-collapse density.

These results are relatively insensitive to cosmology; changing from the standard CDM model used by Tegmark *et al.* to our preferred ΛCDM model results in

[12] Note that there is a typographical error in the equation as given in Tegmark *et al.* (1997), which is corrected here.

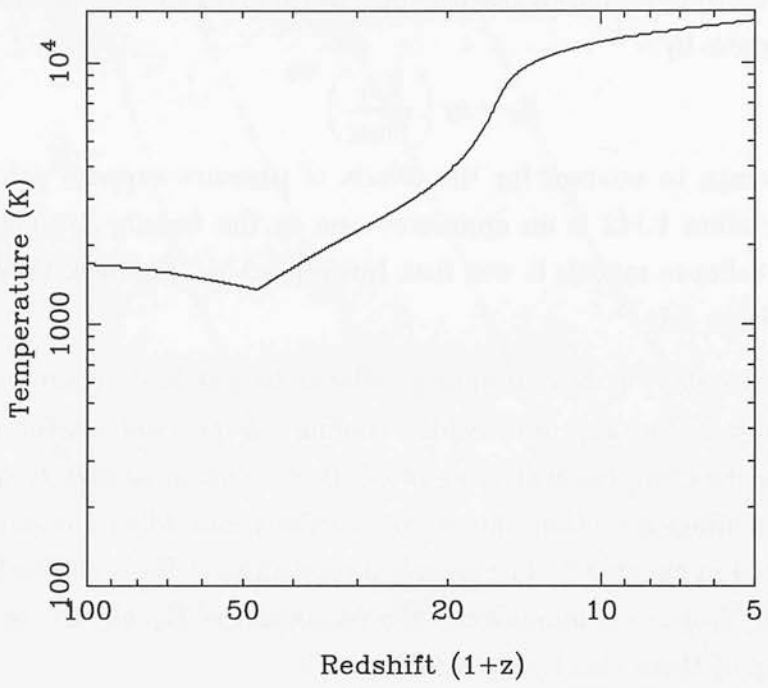

Figure 1.5: $T_{\rm crit}$ as a function of redshift, as computed by Tegmark *et al.* (1997).

values of $T_{\rm crit}$ that are systematically lower by around 10%, but otherwise has no real effect.

Improving the collapse model

A number of improvements can be made to this simple model. To begin with, we can replace the very simplified H_2 chemistry with a more complete model. We can similarly replace the H_2 cooling function used by Tegmark *et al.*, which is based upon that of Hollenbach & Mckee (1979), with a substantially more accurate one from Le Bourlot *et al.* (1999). Full details of the adopted chemistry and cooling model are given in chapter 2.

In figure 1.6 we demonstrate the effects of making these changes. The high redshift evolution of $T_{\rm crit}$ is similar in the two models, but at low redshift more differences are apparent. $T_{\rm crit}$ remains below 10^4 K for a much longer period, but the transition to Lyman-α cooling, when it does come, is far more rapid. These two differences have separate causes. H_2 cooling remains important for a longer period as a result of the choice of cooling function. Tegmark *et al.* considered only the rotational lines of H_2 in their study, implying that the cooling rate at

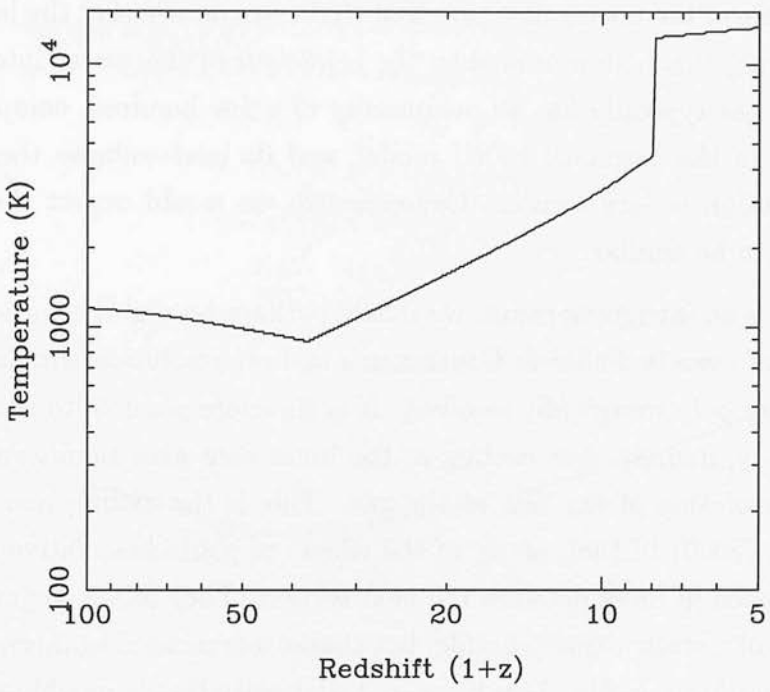

Figure 1.6: $T_{\rm crit}$ as a function of redshift, computed with an improved chemistry and cooling model.

temperatures above 1000 K is significantly underestimated. Consequently, the transition to Lyman-α cooling occurs sooner than it should.

On the other hand, the rapidity of this transition in the improved model is due to the inclusion of the H_2 collisional dissociation reactions (reactions 11-13 in table 2.1). These are negligible below a few thousand K, but become increasingly important at higher temperatures. Eventually, we reach a point at which further temperature increases do not result in an increase in the cooling rate – the rate per molecule continues to increase, but is offset by the smaller H_2 abundance. Once this point is reached, H_2 cooling will no longer be effective, and $T_{\rm crit}$ rises rapidly until Lyman-α cooling begins.

The other area that we might seek to improve is the treatment of the protogalactic density profile – the assumption that protogalaxies have uniform density is simply incorrect, as numerical simulations clearly demonstrate (Abel *et al.*, 2000). However, comparison of the results of the uniform density model with the hydrodynamical simulations of Fuller & Couchman (2000) suggests that the model *already* produces surprisingly accurate results. This is quite unexpected, but consideration of the temperature and density profiles presented in Fuller &

Couchman suggests a possible explanation. The simulations imply that cooling in the inner core of the protogalaxy has little influence on whether the bulk of the gas cools; rather, this is determined by the behaviour of the gas at intermediate radii. [13] This gas typically has an overdensity of a few hundred, comparable to that assumed in the Tegmark *et al.* model, and its post-collapse thermal and chemical evolution is very similar. Consequently, we would expect the inferred values of $T_{\rm crit}$ to be similar.

Although this is an intriguing result, we should perhaps be slightly cautious about accepting it, as even in Fuller & Couchman's highest resolution simulations the core regions are only marginally resolved. It is therefore sensible to consider the other possibility, namely that cooling in the inner core *does* significantly affect the thermal evolution of the rest of the gas. This is the assumption made by Haiman *et al.* (2000) in their study of the effects of photodissociative feedback, which is discussed in more detail in the next section. They follow Tegmark *et al.* in making use of a static density profile, but choose a truncated isothermal sphere rather than a uniform profile. This has a central density that is roughly a hundred times greater than that of the uniform sphere, and this choice corresponds far more closely to the density profiles actually seen in hydrodynamical simulations.

In Haiman *et al.*'s model, the efficiency of cooling is determined by the behaviour of the gas at the edge of the central core. Although they use a different cooling criterion to that of Tegmark *et al.*, it is a simple matter to incorporate their choice of density profile into the latter's model. The resulting evolution of $T_{\rm crit}$ is plotted in figure 1.7. We see that the change to the density profile has a profound effect upon the evolution of $T_{\rm crit}$, which now remains less than 1000 K for the whole period analysed.

Several other groups have also performed simulations of protogalactic formation:

- Haiman *et al.* (1996a) used a spherically symmetric, Lagrangian hydrodynamics code to investigate whether H_2 cooling can significantly affect the thermal evolution of protogalaxies prior to virialization. They also obtained values for $T_{\rm crit}$, finding that $T_{\rm crit} \sim 100$K, with little dependence on redshift. However, this result was obtained using the Lepp & Shull (1983) cooling function, which overestimates the cooling rate at low temperatures by an

[13] Around 10 – 20 pc for the protogalaxies analyzed in figures 8–11 of Fuller & Couchman (2000).

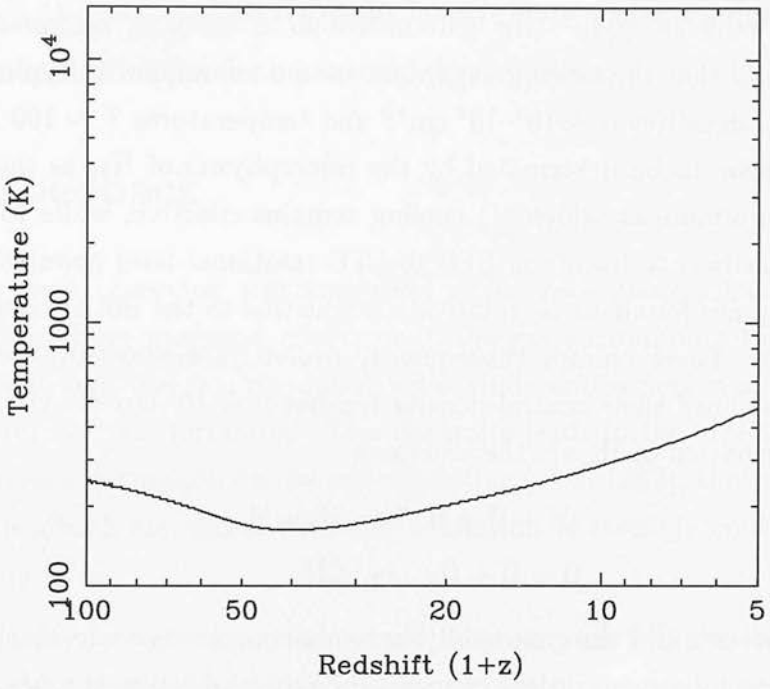

Figure 1.7: The evolution of $T_{\rm crit}$ with redshift for a truncated isothermal sphere density profile.

order of magnitude, and is consequently unreliable.

- Abel *et al.* (2000) and Bromm *et al.* (2002) use three-dimensional hydrodynamical simulations to address the collapse and fragmentation of primordial protogalaxies. Unfortunately, the study of fragmentation requires both high resolution and large dynamical range, leading to a lengthy computational time and limiting the number of cases that can be studied. Since they are interested in protogalaxies that can cool efficiently, Bromm *et al.* (2002) use the Tegmark *et al.* results to guide the choice of suitable initial conditions; they do not determine $T_{\rm crit}$ themselves. By contrast, Abel *et al.* (2000) study a protogalaxy that forms naturally in cosmological hydrodynamics simulation. Their work suggests that $T_{\rm crit} \lesssim 1000$ K at $z \sim 20$, but with essentially only one data point it is difficult to say more than this.

Although not particularly helpful in determining $T_{\rm crit}$, the simulations of Abel *et al.* (2000) and Bromm *et al.* (2002) do give us an insight into the subsequent evolution of the gas inside cooling protogalaxies. Abel *et al.* find that a density profile develops that is approximately isothermal if spherically averaged, but it

is clear both from their figure 2, and from figures 13–23 in Bromm *et al.* (2002), that this averaging masks a great deal of substructure.

Both groups find that the cooling gas fragments into clumps with typical masses $M \sim 1000\ \mathrm{M}_\odot$, densities $n \sim 10^3$–$10^4\ \mathrm{cm}^{-3}$ and temperatures $T \sim 100$ K. These conditions appear to be determined by the microphysics of H_2, as the temperature is the minimum at which H_2 cooling remains effective, while the density marks the transition to from non-LTE to LTE rotational level populations. Accordingly, they are found to be relatively insensitive to the initial conditions of the simulation. These clumps subsequently evolve quasi-statically (with occasional mergers) until their central density reaches $n \sim 10^8\ \mathrm{cm}^{-3}$. At this point, three-body formation of H_2 via the reactions

$$\mathrm{H} + \mathrm{H} + \mathrm{H} \rightarrow \mathrm{H}_2 + \mathrm{H} \tag{1.145}$$

$$\mathrm{H} + \mathrm{H} + \mathrm{H}_2 \rightarrow 2\mathrm{H}_2. \tag{1.146}$$

becomes important, and the gas rapidly becomes completely molecular, boosting the cooling rate by several orders of magnitude and initiating the next stage of collapse.

To follow the subsequent collapse of these clumps towards stellar densities is beyond the capabilities of the simulations presented in Bromm *et al.* (2002) and Abel *et al.* (2000). The adaptive mesh code used in the latter paper has, however, been adapted to study this (Abel *et al.*, 2001). Clumps are found to continue to collapse without any sign of further fragmentation, suggesting that the end point of the collapse is likely to be a massive star. This is in reasonable agreement with the results of Omukai & Nishi (1998), who simulated collapse to stellar densities with a spherically symmetric hydrocode, taking into account radiative transfer in the H_2 lines. [14] Omukai & Nishi find that the initial dense protostellar core is small, but with a rapid accretion rate, and given time we would expect it to accrete much of the surrounding gas in the clump. However, we know from studies of local star formation that stellar outflows play an important role in determining the final stellar mass and this is likely to be true for primordial star formation also. Thus, while there are indications that the first stars were unusually large, the question is far from settled.

To summarize, we have a reasonable understanding of galaxy formation, albeit one that leaves a number of questions still unanswered, up to the point at which

[14] The Abel *et al.* (2001) simulation assumes that the gas remains optically thin.

the first stars form. However, as we shall see in the next section, we can also make substantial progress in understanding what happens *after* the first stars form, provided we make some reasonable assumptions.

1.4 Feedback

We know from observing star formation in our own galaxy that newly-formed massive stars have profound effects upon the gas surrounding them; they photoionize and heat the gas, disrupt it with winds and supernovae, and generally suppress further star formation. This naturally leads to the idea that, in certain conditions, star formation can be self-regulating (Cox, 1983); that there is enough negative feedback associated with star formation to strongly control the rate at which stars form.

It is obviously important to understand whether this will also be true at high redshifts. Moreover, although the precise details of any feedback will obviously depend upon unknowns such as the star formation rate and primordial IMF, I hope to show that the framework within which we can assess its effects is generally independent of these details. Thus, the results presented in chapters 3 and 4 depend upon the assumptions that I make concerning primordial star formation; should these assumptions change, in the light of observational data or better theoretical modelling, it will be a relatively simple matter to recompute the results.

The various feedback effects arising from star formation can be divided into two main kinds. The first kind is what we might call kinetic feedback [15] – the input into the interstellar medium (ISM) of thermal and kinetic energy via stellar winds or supernovae. The second kind of feedback is generally termed radiative feedback, and involves the input of energy or destruction of coolants by radiation.

I do not consider the effects of kinetic feedback in this thesis. In the case of stellar winds, this is easily justified; work by Kudritzki (2000) and Hubeny *et al.* (2000) demonstrates that their effects are likely to be small.

Supernovae, on the other hand, will clearly have a very large influence on their host protogalaxy. For instance, the binding energy of a 10^6 $M_\odot$ protogalaxy forming at a redshift of 20 is of the order of 10^{50} erg, which is significantly smaller

[15] This is also known as stellar feedback (Ciardi *et al.*, 2000a).

than the 10^{51} erg liberated in a typical supernova explosion. Additionally, both analytical modelling (Ferrara & Tolstoy, 2000) and hydrodynamical simulations (Mac Low & Ferrara, 1999) demonstrate that only a few supernovae are necessary to completely disrupt a small protogalaxy, and blow all of its gas out into the intergalactic medium.

However, these conclusions depend to some extent upon the assumptions made concerning the small-scale structure of the gas (see, e.g. Strickland & Stevens, 2000). In particular, if supernovae explode in very dense surroundings ($n > 10^5$ cm^{-3}), then they can reach their radiative cooling phase in only a few years (Terlevich *et al.*, 1992). In this case, the fraction of the initial energy remaining in the form of kinetic energy would be much smaller than is generally supposed, and the amount of mass-loss would be correspondingly reduced.

To properly assess the effect of supernova feedback seems likely to involve high-resolution simulation of the supernova and its surroundings, and, while efforts are being made in this direction (Abel, private communication), it remains a challenging numerical problem that is outside the scope of this thesis.

One firm conclusion that we *can* draw about supernova feedback, however, is its time of onset. We know that it must be delayed for several million years after the formation of the first massive stars, simply because it takes the stars that length of time to exhaust their fuel. In cosmological terms, this delay is negligible; even at $z \sim 25$, it is a small fraction of the Hubble time. Within the protogalaxy, however, this delay is far from negligible; it is comparable to the dynamical timescale of the star forming region and it is quite plausible that significant star formation could occur prior to the explosion of the first supernova.

Radiative feedback within the protogalaxy has the potential to act far faster; typical H_2-cooled protogalaxies have sizes of the order of a few hundred parsecs, so the time taken for the radiation to cross them is not an issue.

We know already that radiative feedback can have a substantial effect. For instance, Ricotti & Shull (2000) show that small stellar clusters can lead to the photoionization of a significant fraction of the protogalactic gas, while Omukai & Nishi (1999) show that they will also destroy the bulk of the H_2. The key question, however, is the timescale on which these effects occur. If this is very much longer than the lifetime of a typical massive star, then radiative effects are likely to be unimportant compared to the effects of supernovae. On the other

hand, if it is very much shorter, then radiative effects will dominate. I consider this issue in more detail in chapter 3.

Another important difference between kinetic and radiative feedbacks is in their area of effect. Kinetic feedback generally affects a limited area, being restricted by the speed at which disturbances can propagate through the intergalactic medium (IGM). For example, a galactic wind propagating into the IGM at a velocity of 1000 km s^{-1} at $z = 25$ will travel only $15h^{-1}$ kpc within a Hubble time, barely 0.1% of the horizon size at this epoch. Thus, while outflows may have important *local* effects (Scannapieco *et al.*, 2000), globally their effects are far more limited. Radiative feedback, on the other hand, can be truly global in scope – if absorption is negligible, then light from a given source will freely propagate throughout the Hubble volume, and gas at any given point will be affected by star formation in many different galaxies.

The most widely studied form of global radiative feedback is reionization. We know from observations of high redshift quasars that the IGM is ionized at a redshift of 5.8 (Fan *et al.*, 2000); on the other hand, the existence of peaks in the CMB power spectrum shows that the universe must have been neutral at high redshift, and allows us to place an upper limit of $z_{\rm ion} \lesssim 35$ on the redshift of reionization (Griffiths *et al.*, 1999). The identity of the sources responsible for reionization is not unambiguously determined, but given the sharp fall in the comoving number density of quasars with increasing redshift (Pei, 1995), it seems likely that star-forming galaxies will be the dominant contributors at high redshift.

Cosmological reionization affects the global star formation rate in several ways. To begin with, it reheats the IGM to $T \sim 10^4$ K, thereby increasing the Jeans mass by several orders of magnitude. [16] Secondly, it suppresses cooling in dwarf galaxies, as outlined in Efstathiou (1992). This effect has been investigated by a number of groups (Thoul & Weinberg, 1996; Kepner *et al.*, 1997; Kaellander & Hultman, 1998; Kitayama & Ikeuchi, 2000), and is found to be significant for galaxies with circular velocities less than 40–50 km s^{-1}, corresponding to virial temperatures $T_{\rm vir} \lesssim 10^5$ K.

Finally, it can destroy protogalaxies with virial temperatures below 10^4 K by

[16] In fact, as demonstrated by Ostriker & Gnedin (1996), this reheating generally precedes reionization, as the hard UV photons that are responsible for much of the heating have a larger mean free path than the near-threshold photons responsible for the bulk of the ionization.

photo-evaporating them (Barkana & Loeb, 1999), leading to the possibility that there may be a substantial population of 'dark' galaxies, with little or no associated gas and stars (Trentham *et al.*, 2001).

Before reionization occurs, however, another form of radiative feedback may occur – H_2 photodissociation by soft UV photons. [17] This occurs via a process known as two-step photodissociation: an H_2 molecule absorbs a UV photon, exciting it to the Lyman or Werner electronic state. The excited molecule subsequently decays to the ground state, with some fraction of these decays being to the vibrational continuum and resulting in the dissociation of the molecule. This is discussed in far more detail in chapter 2.

The importance of UV photodissociation in local star formation has been known for many years (Stecher & Williams, 1967; Heiles, 1971) but only relatively recently has its importance in the early universe been recognised (Haiman *et al.*, 1997; Ciardi *et al.*, 2000b). The most self-consistent treatment of its effects is that of Haiman *et al.* (2000). They show that given some reasonable assumptions concerning star formation efficiency and the primordial IMF, a global soft UV background inevitably develops prior to reionization. This background destroys protogalactic H_2 and suppresses cooling in low-mass halos.

Taken at face value, their results suggest that star formation in H_2-cooled protogalaxies is unimportant; H_2 cooling is suppressed long before such galaxies become common. If true, this is an important result – it is far easier for ionizing photons to escape from these small galaxies than from their larger, Lyman-α cooled counterparts (Ricotti & Shull, 2000) and they are also far more likely to be disrupted by kinetic feedback (Mac Low & Ferrara, 1999). This, together with their larger number density compared to more massive galaxies, implies that they may be the most significant contributors to the reionization and metal enrichment of the IGM *if a substantial number are able to form stars.* Conversely, if photodissociation suppresses star formation in all but a few rare cases, then they will have a negligible effect and may be safely ignored. In view of their potential importance, we should try to ensure that the conclusions reached by Haiman *et al.* (2000) are water-tight.

However, Haiman *et al.* themselves identify the existence of a possible loophole in their argument. If an X-ray background is present, it will raise the fractional

[17] Here and elsewhere, 'soft' UV photons are those below the Lyman limit.

ionization of the protogalactic gas. The increased free electron abundance will in turn catalyze the formation of H_2, thereby promoting H_2 cooling. Haiman *et al.* demonstrate that if this background is sufficiently strong, then the extra H_2 formed can offset the effects of photodissociation, so that no negative feedback occurs.

Haiman *et al.* assume that any X-ray background will be due to an early generation of quasars (or 'mini-quasars') located in dwarf galaxies. They adopt an extremely simple toy model of the background generated by such a population: the X-ray background is taken to be a power-law, $L_\nu \propto \nu^{-1}$, modulated by absorption in the IGM corresponding to a hydrogen column density of 10^{22} cm^{-2}. The normalization is fixed by a parameter f_x, which is the ratio of the (unabsorbed) flux density at the Lyman limit to the flux density in the Lyman-Werner bands. They find that one requires $f_x \gtrsim 0.1$ to overcome negative feedback. Given the uncertainty that surrounds the origin of quasars, it is far from clear that enough exist in the early universe to produce a background of this size. Indeed, there is some observational evidence against this model (Haiman *et al.*, 1999).

However, it is not the only possibility. We know from observations of local galaxies that star formation itself leads to the production of X-rays; the X-ray luminosity of star-forming galaxies is closely linked to their star-formation rate (Moran *et al.*, 1999). There seems no reason why this should not also be true at high redshift; indeed, Oh (2001) argues that many of the possible production mechanisms for the X-rays should be *more* effective at high redshift. It is thus quite conceivable that primordial star-formation will be accompanied by enough X-ray production to prevent photodissociation feedback from occurring; or, at the very least, to substantially mitigate its effects. A major aim of this thesis is to investigate the plausibility of this suggestion.

Accordingly, in chapter 4 I investigate the effects of various different X-ray sources models on protogalactic cooling. Using a strategy similar to that of Haiman *et al.* (2000), I attempt to calculate the evolution of $T_{\rm crit}$ and of the X-ray and Lyman-Werner backgrounds self-consistently. I also examine the dependence of the results on the various assumptions made in my simulations; in particular, on the approximations that must be made in order to model H_2 photodissociation.

1.5 Thesis outline

In chapter 2, I present a simplified chemical model of primordial gas. This builds on the work of Abel *et al.* (1997), but has been extended in order to correctly treat the chemistry of optically thick gas. I also discuss in detail the approximations that underly this simple model and the accuracy with which the reaction rates are known. Finally, I briefly discuss the various different ways in which primordial gas cools.

In chapter 3, I look at the effects of radiative feedback on small scales, *i.e.* inside individual protogalaxies. Using a simple protogalactic density profile, I examine the growth and final fate of H II regions within the protogalaxy and assess the overall importance of photoionization. Similarly, I examine the importance of photodissociation, using an approximation which properly treats the effects of H_2 self-shielding. I also investigate the effects of ionizing and dissociating radiation on dense clumps of gas – the likely site of future star formation.

Finally, in chapter 4 I look at the large-scale effects of radiative feedback. I confirm the negative feedback effect found by Haiman *et al.* (2000), using a different computational method, but show that if one also includes the effects of the X-rays associated with star formation then this feedback effect is substantially reduced.

Chapter 2

Modelling primordial gas chemistry

2.1 Introduction

Essential to the understanding of the formation and evolution of the first protogalaxies is an understanding of their chemical evolution. In particular, we need to be able to understand the chemistry of molecular hydrogen, H_2, which dominates cooling at low temperatures. To do this requires an accurate chemical model. In this chapter I discuss how such a model is constructed.

We start by recognising that, prior to star formation, the only elements present in the gas will be those produced by primordial nucleosynthesis. Recent determinations are summarized in table 1.1; essentially, primordial gas is a mixture of hydrogen and helium, with small traces of D, ^{3}He and Li. The number of possible reactions is thus very much smaller than in the local ISM. Despite this, a great number of reactions are still possible, particularly once we include the effects of radiation.

If we are to avoid getting bogged down in detail, we clearly need to make some simplifications. The key to doing this is the fact that we are not really interested in the primordial chemistry *per se*; rather, we are interested in its effects on the thermal evolution of primordial gas. Elements of the chemistry which do not affect this evolution, directly or indirectly, can be ignored.

Building a model along these lines from scratch would be a time-consuming task. Fortunately, much work in this area has already been done, in particular by Abel

et al. (1997) and Galli & Palla (1998). Both groups present simplified chemical models designed with particular applications in mind. The Abel *et al.* model is designed for studying galaxy formation in large-scale hydrodynamical simulations and implicitly assumes that the gas will be optically thin; the Galli & Palla model, on the other hand, was designed primarily for studying the chemistry of the primordial IGM and assumes that any source of radiation will be thermal.

The main drawback of both models is their treatment of the photochemistry. Neither model is suited to studying gas which is optically thick, and which may be illuminated by a fairly arbitrary spectrum of background radiation. Thus, while the models provide a substantial framework on which to build, they must be supplemented by an improved photochemical treatment.

As the Abel *et al.* model is closer to what is required than the Galli & Palla model, I take the former as a basis. The particular changes which are required to the model are the inclusion of the effects of secondary ionization and H_2 self-shielding. Additionally, a more accurate treatment of helium recombination is also required. I have also taken the opportunity to add in a few reactions overlooked by Abel *et al.*, and to update various of the reaction rates in the light of new data.

The reactions included in the chemical model are set out in tables 2.1 and 2.1, along with analytical fits to the rate coefficients (or radiative cross-sections), plus a reference to the primary source of the data. [1] In view of the similarity of much of the model to that of Abel *et al.*, I have adopted their reaction numbering scheme, with appropriate extensions.

In section 2.2, I discuss the approximations upon which the model is based, and attempt to justify them. The reactions themselves are discussed in sections 2.3 and 2.4. In section 2.5, I briefly discuss the various processes which heat and cool the gas while in section 2.6 I summarize my results.

[1] Where the fitting formula is from a source other than the primary reference, this is indicated in the text.

No.	Reaction	Rate coefficient ($\mathrm{cm^3\,s^{-1}}$)	Reference
1.	$\mathrm{H + e^- \rightarrow H^+ + 2e^-}$	$k_1 = \exp[-32.71396786 + 13.536556 \ln T_e - 5.73932875 (\ln T_e)^2 + 1.56315498 \times 10^0 (\ln T_e)^3 - 2.877056 \times 10^{-1} (\ln T_e)^4 + 3.48255977 \times 10^{-2} (\ln T_e)^5 - 2.63197617 \times 10^{-3} (\ln T_e)^6 + 1.11954395 \times 10^{-4} (\ln T_e)^7 - 2.03914985 \times 10^{-6} (\ln T_e)^8]$	Janev *et al.* (1987)
2.	$\mathrm{H^+ + e^- \rightarrow H + \gamma}$	Case A: $k_{2a} = 1.269 \times 10^{-13} \frac{(315614/T)^{1.503}}{[1.0+(604625/T)^{0.470}]^{1.923}}$ Case B: $k_{2b} = 2.753 \times 10^{-14} \frac{(315614/T)^{1.500}}{[1.0+(115188/T)^{0.407}]^{2.242}}$	Ferland *et al.* (1992)
3.	$\mathrm{He + e^- \rightarrow He^+ + 2e^-}$	$k_3 = \exp[-44.09864886 + 23.91596563 \ln T_e - 10.7532302 (\ln T_e)^2 + 3.05803875 \times 10^0 (\ln T_e)^3 - 5.6851189 \times 10^{-1} (\ln T_e)^4 + 6.79539123 \times 10^{-2} (\ln T_e)^5 - 5.00905610 \times 10^{-3} (\ln T_e)^6 + 2.06723616 \times 10^{-4} (\ln T_e)^7 - 3.64916141 \times 10^{-6} (\ln T_e)^8]$	Janev *et al.* (1987)
4.	$\mathrm{He^+ + e^- \rightarrow He^{++} + 2e^-}$	$k_4 = \exp[-68.71040990 + 43.93347633 \ln T_e - 18.4806699 (\ln T_e)^2 + 4.70162649 \times 10^0 (\ln T_e)^3 - 7.6924663 \times 10^{-1} (\ln T_e)^4 + 8.11304200 \times 10^{-2} (\ln T_e)^5 - 5.32402063 \times 10^{-3} (\ln T_e)^6 + 1.97570531 \times 10^{-4} (\ln T_e)^7 - 3.16558106 \times 10^{-6} (\ln T_e)^8]$	Abel *et al.* (1997)

continued on next page

No.	Reaction	Rate coefficient ($\mathrm{cm^3\,s^{-1}}$)	Reference
5.	$\mathrm{He^+ + e^- \rightarrow He + \gamma}$	Case A: $k_{5a} = 10^{-11}T^{-1/2}[12.72 - 1.615\log(T) - 0.3162\log(T)^2 + 4.930\times10^{-2}\log(T)^3]$	Hummer & Storey (1998)
		Case B: $k_{5b} = 10^{-11}T^{-1/2}[11.19 - 1.676\log(T) - 0.2852\log(T)^2 + 4.433\times10^{-2}\log(T)^3]$	
		Dielectronic: $k_{5di} = 1.9\times10^{-3}T^{-3/2}\exp\left(\frac{-4.7\times10^5}{T}\right)\left[1 + 0.3\exp\left(\frac{-9.4\times10^4}{T}\right)\right]$	Aldrovandi & Pequignot (1973)
6.	$\mathrm{He^{++} + e^- \rightarrow He^+ + \gamma}$	Case A: $k_{6a} = 2\times k_{2a}(T/4)$	Ferland *et al.* (1992)
		Case B: $k_{6b} = 2\times k_{2b}(T/4)$	

continued on next page

No.	Reaction	Rate coefficient ($\mathrm{cm^3\,s^{-1}}$)		Reference
7.	$\mathrm{H + e^- \rightarrow H^- + \gamma}$	$k_7 = \mathrm{dex}[-17.845 + 0.762\log(T) + 0.1523\log(T)^2$ $-3.274\times10^{-2}\log(T)^3]$	$T < 6000$ K	Wishart (1979)
		$k_7 = \mathrm{dex}[-16.420 + 0.1998\log(T)^2 - 5.447\times10^{-3}\log(T)^4$ $+4.0415\times10^{-5}\log(T)^6]$	$T > 6000$ K	
8.	$\mathrm{H^- + H \rightarrow H_2 + e^-}$	$k_8 = \begin{cases} 1.5\times10^{-9} \\ 4.0\times10^{-9}T^{-0.17} \end{cases}$	$T < 300$K $T > 300$K	Launay *et al.* (1991)
9.	$\mathrm{H + H^+ \rightarrow H_2^+ + \gamma}$	$k_9 = \mathrm{dex}[-19.38 - 1.523\log \mathrm{T} + 1.118\log(\mathrm{T})^2 - 0.1269\log(\mathrm{T})^3]$		Ramaker & Peek (1976)
10.	$\mathrm{H_2^+ + H \rightarrow H_2 + H^+}$	$k_{10} = 6.4\times10^{-10}$		Karpas *et al.* (1979)
11.	$\mathrm{H_2 + H^+ \rightarrow H_2^+ + H}$	$k_{11} = \begin{cases} 3.0\times10^{-10}\exp\left(\frac{-21050}{T}\right) \\ 1.5\times10^{-10}\exp\left(\frac{-14000}{T}\right) \end{cases}$	$T < 10^4$K $T > 10^4$K	Holliday *et al.* (1971)
12.	$\mathrm{H_2 + e^- \rightarrow 2H + e^-}$	$k_{12} = 3.73\times10^{-9}T^{0.1121}\exp\left(\frac{-99430}{T}\right)$		Stibbe & Tennyson (1999)

continued on next page

No.	Reaction	Rate coefficient ($cm^3\ s^{-1}$)	Reference
13.	$H_2 + H \rightarrow 3H$	$k_{13} = k_{cd} + k_{dt}$	Martin *et al.* (1996)
		$k_{cd} = \mathrm{dex}\,[128.8953 - 53.91334\log(T) + 5.315517\log(T)^2 - 19.73427\log(1 + 16780.95/T) - 25786.11/T]$	
		$k_{dt} = \mathrm{dex}\,[60.34928 - 27.43096\log(T) + 2.676150\log(T)^2 - 11.28215\log(1 + 14254.55/T) - 23125.2/T]$	
14.	$H^- + e^- \rightarrow H + 2e^-$	$k_{14} = \exp[-18.01849334 + 2.3608522\ln T_e - 0.28274430(\ln T_e)^2 + 1.62331664 \times 10^{-2}(\ln T_e)^3 - 3.36501203 \times 10^{-2}(\ln T_e)^4 + 1.17832978 \times 10^{-2}(\ln T_e)^5 - 1.65619470 \times 10^{-3}(\ln T_e)^6 + 1.06827520 \times 10^{-4}(\ln T_e)^7 - 2.63128581 \times 10^{-6}(\ln T_e)^8]$	Janev *et al.* (1987)
15.	$H^- + H \rightarrow 2H + e^-$	$k_{15} = 2.5634 \times 10^{-9} T_e^{1.78186}$ $T_e < 0.1$ eV	Janev *et al.* (1987)
		$k_{15} = \exp[-20.372609 + 1.13944933\ln T_e - 0.14210135(\ln T_e)^2 + 8.4644554 \times 10^{-3}(\ln T_e)^3 - 1.4327641 \times 10^{-3}(\ln T_e)^4 + 2.0122503 \times 10^{-4}(\ln T_e)^5 + 8.6639632 \times 10^{-5}(\ln T_e)^6 - 2.5850097 \times 10^{-5}(\ln T_e)^7 + 2.4555012 \times 10^{-6}(\ln T_e)^8 - 8.0683825 \times 10^{-8}(\ln T_e)^9]$ $T_e > 0.1$ eV	

continued on next page

No.	Reaction	Rate coefficient ($cm^3\ s^{-1}$)		Reference
16.	$H^- + H^+ \rightarrow 2H$	$k_{16} = 5.7 \times 10^{-6} T^{-0.5} + 6.3 \times 10^{-8} - 9.2 \times 10^{-11} T^{0.5}$ $+4.4 \times 10^{-13} T$		Moseley *et al.* (1970)
17.	$H^- + H^+ \rightarrow H_2^+ + e^-$	$k_{17} = \begin{cases} 6.9 \times 10^{-9} T^{-0.35} \\ 9.6 \times 10^{-7} T^{-0.9} \end{cases}$	$T < 8000$ K $T > 8000$ K	Poulaert *et al.* (1978)
18.	$H_2^+ + e^- \rightarrow 2H$	$k_{18} = \begin{cases} 1.0 \times 10^{-8} \\ 1.32 \times 10^{-6} T^{-0.76} \end{cases}$	$T < 617$K $T > 617$K	Schneider *et al.* (1994)
29.	$H^- + He \rightarrow H + He + e^-$	$k_{29} = 4.1 \times 10^{-17} T^2 \exp\left(\frac{-19870}{T}\right)$		Huq *et al.* (1982)
30.	$He^+ + H \rightarrow He + H^+ + \gamma$	$k_{30} = 1.20 \times 10^{-15} \left(\frac{T}{300}\right)^{0.25}$		Zygelman *et al.* (1989)
31.	$He + H^+ \rightarrow He^+ + H$	$k_{31} = \begin{cases} 1.26 \times 10^{-9} T^{-0.75} \exp\left(\frac{-127500}{T}\right) \\ 4 \times 10^{-37} T^{4.74} \end{cases}$	$T < 10^4$ K $T > 10^4$ K	Kimura *et al.* (1993)

Table 2.1: Analytical fits to the rate coefficients for all of the purely collisional reactions in the chemical model, plus references to the source of the data. Further details of all of the reactions are given in section 2.3.

No.	Reaction	Cross-section (cm^{-2})		Reference
20.	$\mathrm{H} + \gamma \rightarrow \mathrm{H}^+ + \mathrm{e}^-$	$\sigma_{20} = 6.3 \times 10^{-18} \eta^4 \exp(4 - 4\epsilon^{-1} \arctan\ \epsilon)[1 - \exp(-2\pi/\epsilon)]^{-1}$		Osterbrock (1989)
21.	$\mathrm{He} + \gamma \rightarrow \mathrm{He}^+ + \mathrm{e}^-$	$\sigma_{21} = 3.1451 \times 10^{-16} \eta^{7/2} [1 - 4.7416\eta^{1/2} + 14.82\eta - 30.8678\eta^{3/2}$ $+37.3584\eta^2 - 23.4585\eta^{5/2} + 5.9133\eta^3]$		Yan *et al.* (1998)
22.	$\mathrm{He}^+ + \gamma \rightarrow \mathrm{He}^{++} + \mathrm{e}^-$	$\sigma_{22} = 1.58 \times 10^{-18} \eta^4 \exp(4 - 4\epsilon^{-1} \arctan\ \epsilon)[1 - \exp(-2\pi/\epsilon)]^{-1}$		Osterbrock (1989)
23.	$\mathrm{H}^- + \gamma \rightarrow \mathrm{H} + \mathrm{e}^-$	$\sigma_{23} = 7.928 \times 10^5 (\nu - \nu_{\mathrm{th}})^{3/2} \nu^{-3}$		de Jong (1972)
		$\sigma_{24} = 9.548 \times 10^{-17} \eta^{-1} - 9.4 \times 10^{-17}$	$h\nu < 16.50$ eV	
		$\sigma_{24} = 2.16 \times 10^{-17} \eta^{-1} - 1.48 \times 10^{-17}$	$16.50 < h\nu < 17.70$ eV	O'Neil & Reinhardt (1978)
		$\sigma_{24} = 1.51 \times 10^{-17} \eta^{-2.71}$	$17.70 < h\nu < 30.0$ eV	
24a.	$\mathrm{H}_2 + \gamma \rightarrow \mathrm{H}_2^+ + \mathrm{e}^-$			
		$\sigma_{24} = 10^{-18} [0.664 - 11.768\eta + 78.118\eta^2 - 231.339\eta^3$		
24b.	$\mathrm{H}_2 + \gamma \rightarrow \mathrm{H} + \mathrm{H}^+ + \mathrm{e}^-$	$+368.053\eta^4 - 189.953\eta^5]$	$30 < h\nu < 85$ eV	Wilms *et al.* (2000)
		$\sigma_{24} = 10^{-16} \eta^{7/2} [1 - 2.003\eta^{1/2} - 4.806\eta + 50.577\eta^{3/2}$		
		$-171.044\eta^2 + 231.608\eta^{5/2} - 81.885\eta^3]$	$h\nu > 85$ eV	Yan *et al.* (1998)

continued on next page

No.	Reaction	Cross-section (cm^{-2})	Reference
25.	$H_2^+ + \gamma \rightarrow H + H^+$	$\sigma_{25} = \mathrm{dex}\,[-40.97 + 15.9795\eta^{-1} - 3.53934\eta^{-2} + 0.2581155\eta^{-3}]$ $2.65 < h\nu < 11.27$ eV $\sigma_{25} = \mathrm{dex}\,[-30.26 + 7.3935\eta^{-1} - 1.29214\eta^{-2} + 6.5785 \times 10^{-2}\eta^{-3}]$ $11.27 < h\nu < 21.0$ eV	Dunn (1968)
26.	$H_2^+ + \gamma \rightarrow 2H^+ + e^-$	$\sigma_{26} = \mathrm{dex}\,[-16.926 - 1.2774\eta^{-1} + 0.20142\eta^{-2} + 1.14615 \times 10^{-2}\eta^{-3}]$ $30.0 < h\nu < 90.0$ eV	Bates & Öpik (1968)
27.	$H_2 + \gamma \rightarrow H_2^* \rightarrow 2H$	See text	–

Table 2.2: Analytical fits to the radiative cross-sections of the included photochemical reactions. In these fits, $\eta \equiv \nu_{\mathrm{th}}/\nu$; the various threshold energies are given in table 2.3. Any other special symbols are defined in the text.

No.	Reaction	Energy threshold (eV)	Reference
20.	$H + \gamma \rightarrow H^+ + e^-$	13.6	Osterbrock (1989)
21.	$He + \gamma \rightarrow He^+ + e^-$	24.6	Yan *et al.* (1998)
22.	$He^+ + \gamma \rightarrow He^{++} + e^-$	54.4	Osterbrock (1989)
23.	$H^- + \gamma \rightarrow H + e^-$	0.755	de Jong (1972)
24a.	$H_2 + \gamma \rightarrow H_2^+ + e^-$	15.4	O'Neil & Reinhardt (1978)
24b.	$H_2 + \gamma \rightarrow H + H^+ + e^-$	18.08	Samson & Haddad (1994)
25.	$H_2^+ + \gamma \rightarrow H + H^+$	2.65	Dunn (1968)
26.	$H_2^+ + \gamma \rightarrow 2H^+ + e^-$	30.0	Bates & Öpik (1968)
27.	$H_2 + \gamma \rightarrow H_2^* \rightarrow 2H$	See text	—

Table 2.3: The threshold energies (in eV) of the various photochemical reactions.

2.2 Approximations

Approximation 1: No D, Li or ^{3}He

These elements have small fractional abundances and play no significant role in determining the H_2 abundance. Deuterium and lithium may, however, affect the cooling rate at low temperatures, where H_2 cooling is ineffective. A complete thermal model should include these coolants. In practice, however, the molecular cooling rate at these temperatures is generally too small to be of practical interest.

A simple order-of-magnitude estimate of the HD abundance required for efficient cooling can be obtained by equating the HD cooling timescale with the dynamical timescale. In a gas of total number density n and temperature T, these timescales are given by

$$t_{\rm cool} = \frac{nkT}{\Lambda_{\rm HD} n^2 f_{\rm HD}}, \tag{2.1}$$

(where $f_{\rm HD}$ is the fractional abundance of HD and $\Lambda_{\rm HD}$ is the HD cooling function) and

$$t_{\rm dyn} \simeq (G\rho)^{-1/2}$$

$$\simeq 10^8 n^{-1/2} \text{ yr}. \tag{2.2}$$

Using the HD cooling function from Galli & Palla (1998), we obtain the abundance required at 100 K

$$f_{\text{HD}} = 2.8 \times 10^{-5} n^{-1/2}. \tag{2.3}$$

As the typical HD/H_2 ratio is of the order of 10^{-3} (Galli & Palla, 1998), this implies a corresponding H_2 abundance of $3 \times 10^{-2} n^{-1/2}$. This is larger than we would expect to find in low density gas – fractional abundances of 10^{-3} to 10^{-4} are the norm – and thus HD cooling will generally only be significant at high density. This is born out by the work of Bromm *et al.* (2002), who find that the inclusion of HD cooling in their simulations of protogalactic collapse leads to differences in behaviour only when one approaches the scale of the dense ($n > 10^3$ cm^{-3}) sub-galactic scale clumps that form at the centre of their simulated protogalaxies.

A similar analysis can be performed for LiH, but in this case the discrepancy between the required and actual abundance is even larger.

Approximation 2: No excited states

The second assumption that we make is that all atoms and molecules are to be found in their ground states; in other words, that we do not need to explicitly calculate their level populations.

For atoms, the accuracy of this assumption is easy to verify. The largest excited population is that of the metastable 2^3S state of He and even this is significantly populated only for electron densities greater than 10^3 cm^{-3} (Osterbrock, 1989). For other states of hydrogen and helium the required density is orders of magnitude larger.

For H^-, the analysis is also simple. The only stable excited states known lie more than 9.5 eV above the ground state (Snow, 1973). This is far larger than the energy required to dissociate H^-, and hence most collisions result in dissociation rather than excitation.

For H_2, the analysis is much more complicated. H_2 molecules have small transition probabilities and closely spaced rotational and vibrational energy levels, which are far easier to populate than excited atomic states. If we neglect radiative pumping and the effects of H_2 formation and destruction, then the rotational

and vibrational level populations will relax fairly quickly towards their statistical equilibrium values, given by

$$n_i = \frac{\sum_{j \neq i} n_j \left(A_{ji} + \sum_c q_{ji,c} n_c \right)}{\sum_{j \neq i} \left(A_{ij} + \sum_c q_{ij,c} n_c \right)}, \tag{2.4}$$

where A_{ij} is the spontaneous transition probability from level i to level j (zero if $E_j > E_i$) and $q_{ij,c}$ the collisional transition rate for collisions with a partner c. As the relaxation timescale is less than a year for the vibrational levels, and less than 10^3 yr for the rotational levels, the statistical equilibrium level populations provide a good estimate of the true values.

To solve equation 2.4 for the equilibrium populations, I have used molecular data from a number of sources. Radiative transition rates are taken from Wolniewicz *et al.* (1998), while collisional rates for reactive and non-reactive H–H_2 collisions, He–H_2 collisions and H^+–H_2 collisions come from Le Bourlot *et al.* (1999), Flower & Roueff (1998a), Flower *et al.* (1998) and Gerlich (1990) respectively. Collisions of H_2 with H_2 can be neglected due to the small H_2 abundance. Selected results of these calculations are plotted in figures 2.1 to 2.3.

We see that the vibrational level populations remain negligible for all but the highest temperatures and densities and can be safely ignored. On the other hand, the lower-lying rotational levels develop significant populations even at moderate temperatures and densities, suggesting that our approximation is inaccurate for these levels.

How seriously does this affect our model? Generally, the effects are small. For instance, the energy difference between the lower rotational levels and the ground state is small compared to the molecular binding energy, and consequently we would not expect rotational excitation to significantly affect the collisional dissociation rates. Similarly, the optically thin photodissociation rate is insensitive to rotational excitation. However, as we shall see later, the photodissociation rate in optically thick gas *can* be significantly affected. I discuss this in more detail later.

Adding in the effects of radiative pumping might be thought to make this problem worse, by providing an additional means by which to create excited H_2. In actual fact, it generally simplifies matters. In order for radiative pumping to create a significant level population, it must occur on a timescale comparable to the lifetime of the excited level. However, since $k_{\rm dis} \simeq 0.2 k_{\rm pump}$, this implies that H_2

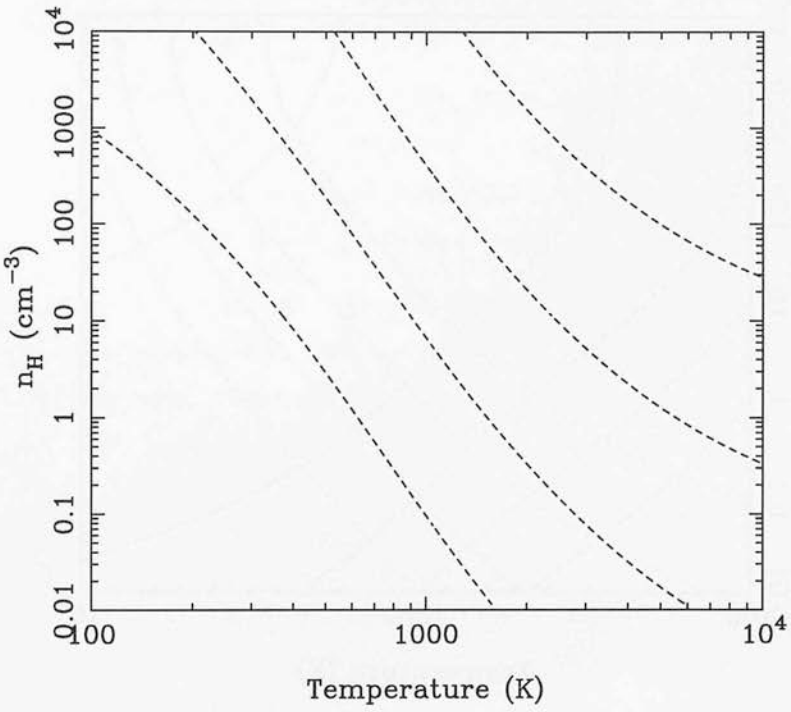

Figure 2.1: Contour plot of the fraction of H_2 molecules in vibrational level $v = 1$, as a function of density and temperature. Contours, from left to right, represent fractions of 10^{-8}, 10^{-6}, 10^{-4} and 10^{-2}. The fractional ionization was 10^{-4}.

dissociation also occurs on a similar timescale. Moreover, since this timescale is generally short (10^3 yr or less) compared to the H_2 formation timescale, radiative pumping significantly affects the level populations only when H_2 is being rapidly destroyed. In such situations, however, we generally do not need to calculate the photodissociation rate to a high degree of accuracy, and so can ignore radiative pumping.

We find similar behaviour in the case of H_2^+ – vibrational levels are populated only at high densities ($n_e \gtrsim 100$ cm^{-3}; Sarpal & Tennyson 1993) or at high redshift ($z \gtrsim 200$), but rotational levels are populated at moderate temperatures and densities. Unlike H_2, however, the self-shielding of H_2 is not significant, and thus the details of the rotational excitation can generally be ignored.

Approximation 3: No three-body processes

At high densities, the three body reactions

$$H + H + H \rightarrow H_2 + H \tag{2.5}$$

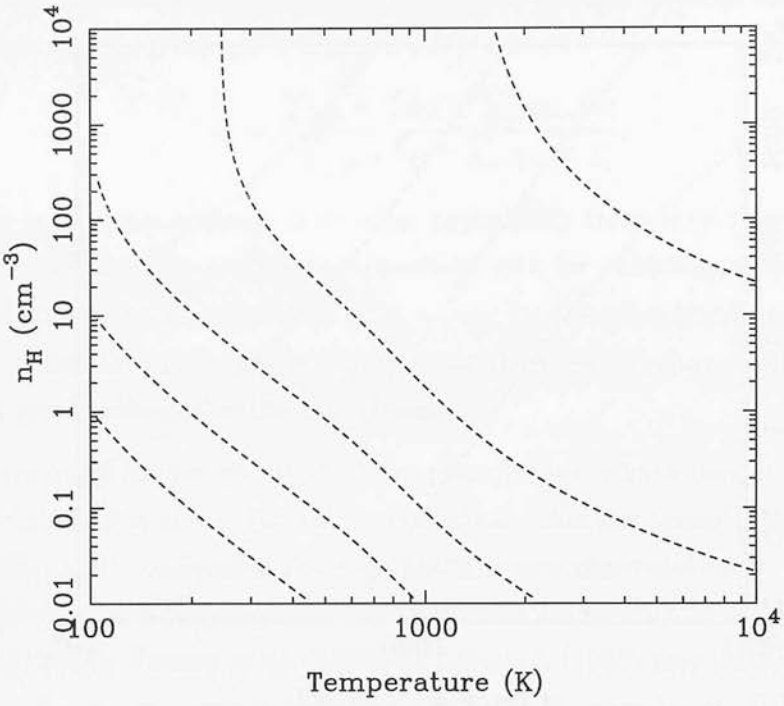

Figure 2.2: Contour plot of the fraction of H_2 molecules in rotational level $J = 2$. Contours, from left to right, represent fractions of 10^{-4}, 10^{-3}, 10^{-2}, and 0.1 (twice). At high densities and temperatures, the fraction decreases as more highly excited states become populated.

and

$$H + H + H_2 \rightarrow 2H_2. \tag{2.6}$$

dominate the H_2 chemistry and are capable of driving the H_2 abundance close to one (Palla *et al.*, 1983). However, the density required for these reactions to become significant is of the order of 10^8 cm^{-3}, which is encountered only at the centre of dense, star-forming clumps.

Approximation 4: Simplified molecular chemistry

Omitting deuterium and lithium removes much of the possible molecular chemistry, while some of that which remains, such as the chemistry of H^- or H_2^+, is essential to the model. However, there are also a number of reactions involving other molecular ions, in particular HeH^+ and H_3^+. These need only be included if they significantly affect the cooling, whether directly, by HeH^+ or H_3^+ acting as coolants, or indirectly, by affecting the H^-, H_2^+ or H_2 abundances.

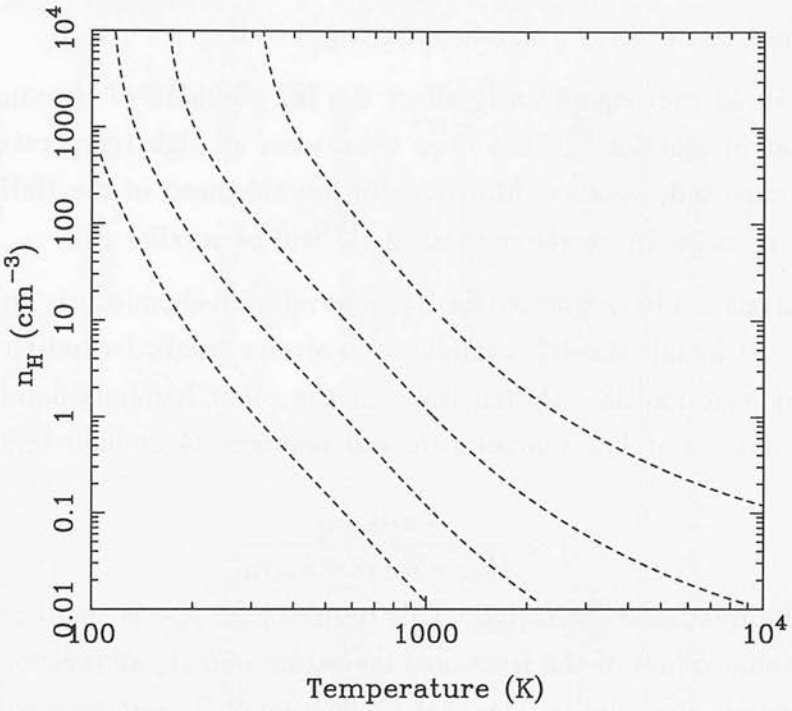

Figure 2.3: As figure 2.2, but for rotational level $J = 3$.

The main collisional reactions involving HeH^+ are listed in table 2.4. Given this set of reactions, it is easy to demonstrate that the HeH^+ abundance will be small – reactions 32 and 34 are very slow, while reaction 33 is limited by a small H_2^+ abundance. Since the neutral hydrogen abundance is generally significantly larger than the fractional ionization or H_2 abundance, destruction of HeH^+ will be dominated by reaction 35. A simple estimate of the equilibrium HeH^+ abundance can thus be constructed. If formation is dominated by reaction 33 then

$$\begin{aligned} f_{HeH^+} &= \frac{k_{33}}{k_{35}} \frac{n_{He}}{n_H} f_{H_2^+} \\ &\leq 10^{-2} f_{H_2^+} \end{aligned} \tag{2.7}$$

for $10^2 < T < 10^4$ K. Alternatively, if reaction 32 dominates we have

$$\begin{aligned} f_{HeH^+} &= \frac{k_{32}}{k_{35}} \frac{n_{He}}{n_H} f_{H^+} \\ &\leq 10^{-11} f_{H^+} \end{aligned} \tag{2.8}$$

over the same temperature range, while if reaction 34 dominates we have

$$\begin{aligned} f_{HeH^+} &= \frac{k_{34}}{k_{35}} f_{He^+} \\ &\leq 10^{-8} f_{He^+}. \end{aligned} \tag{2.9}$$

In all cases, the HeH^+ abundance is negligible. A corollary of this result is that reaction 37 does not destroy a significant amount of H_2.

Could reaction 33 ever significantly affect the H_2^+ abundance? Comparing the rate with that of reaction 10 it is clear that, even at high temperatures, it is an order of magnitude smaller. Moreover, in practice most of the HeH^+ will be destroyed by reaction 35, so the net loss of H_2^+ will be smaller still.

A similar analysis can be performed for H_3^+. The relevant chemistry is summarized in table 2.5. Although the H_2^+ abundance is always small, formation of H_3^+ is dominated by reaction 38. Destruction, on the other hand, is dominated by reactions 41 and 42 at low temperature and reaction 44 at high temperature. Thus

$$f_{H_3^+} \simeq \frac{k_{38} f_{H_2} f_{H_2^+}}{(k_{41} + k_{42})x + k_{44} f_H}, \tag{2.10}$$

where x is the fractional ionization. This implies that $f_{H_3^+}$ is smaller than $f_{H_2^+}$ for all reasonable values of the fractional ionization and H_2 abundance. H_3^+ can thus be neglected. This also implies that production of H_2 and destruction of H^- by H_3^+ can be ignored. Moreover, we note that destruction of H_2 via reaction 38 is negligible compared to the other reactions which destroy H_2. It is thus safe to ignore H_3^+ chemistry.

No.	Reaction	Rate coefficient ($\mathrm{cm^3\,s^{-1}}$)	Reference
32.	$\mathrm{He} + \mathrm{H^+} \rightarrow \mathrm{HeH^+} + \gamma$	$k_{32} = 8.0 \times 10^{-20} \left(\frac{T}{300}\right)^{-0.24} \exp\left(\frac{-T}{4000}\right)$	Juřek *et al.* (1995)
33.	$\mathrm{He} + \mathrm{H_2^+} \rightarrow \mathrm{HeH^+} + \mathrm{H}$	$k_{33} = 3.0 \times 10^{-10} \exp\left(\frac{-6717}{T}\right)$	Black (1978)
34.	$\mathrm{He^+} + \mathrm{H} \rightarrow \mathrm{HeH^+} + \gamma$	$k_{34} = 4.16 \times 10^{-16} \left(\frac{T}{300}\right)^{-0.37} \exp\left(\frac{-T}{87600}\right)$	Kraemer *et al.* (1995)
35.	$\mathrm{HeH^+} + \mathrm{H} \rightarrow \mathrm{He} + \mathrm{H_2^+}$	$k_{35} = 1.04 \times 10^{-9} \left(\frac{T}{300}\right)^{0.13} \exp\left(\frac{-T}{33100}\right)$	Linder *et al.* (1995)
36.	$\mathrm{HeH^+} + \mathrm{e^-} \rightarrow \mathrm{He} + \mathrm{H}$	$k_{36} = 3.0 \times 10^{-8} \left(\frac{T}{300}\right)^{-0.47}$	Guberman (1994)
37.	$\mathrm{HeH^+} + \mathrm{H_2} \rightarrow \mathrm{H_3^+} + \mathrm{He}$	$k_{37} = 1.53 \times 10^{-9} \left(\frac{T}{300}\right)^{0.24} \exp\left(\frac{-T}{14800}\right)$	Linder *et al.* (1995)

Table 2.4: Analytical fits (taken from Stancil *et al.* 1998) to the rate coefficients for the reactions determining the $\mathrm{HeH^+}$ abundance.

No.	Reaction	Rate coefficient ($cm^3\ s^{-1}$)	Reference
38.	$H_2^+ + H_2 \rightarrow H_3^+ + H$	$k_{38} = 2.24 \times 10^{-9} \left(\frac{T}{300}\right)^{0.042} \exp\left(\frac{-T}{46600}\right)$	Linder *et al.* (1995)
39.	$H_2 + H^+ \rightarrow H_3^+ + \gamma$	$k_{39} = 1.0 \times 10^{-20}$	Dalgarno & McDowell (1956)
40.	$H_2^+ + H \rightarrow H_3^+ + \gamma$	$k_{40} = 1.5 \times 10^{-17} \left(\frac{T}{300}\right)^{1.8} \exp\left(\frac{20}{T}\right)$	Dalgarno & McDowell (1956)
41.	$H_3^+ + e^- \rightarrow H_2 + H$	$k_{41} = 4.0 \times 10^{-8} \left(\frac{T}{300}\right)^{-0.5}$	Datz *et al.* (1995)
42.	$H_3^+ + e^- \rightarrow 3H$	$k_{42} = 1.6 \times 10^{-7} \left(\frac{T}{300}\right)^{-0.5}$	Datz *et al.* (1995)
43.	$H_3^+ + H^- \rightarrow H_2 + 2H$	$k_{43} = 2.3 \times 10^{-7} \left(\frac{T}{300}\right)^{-0.5}$	Dalgarno & McDowell (1956)
44.	$H_3^+ + H \rightarrow H_2 + H_2^+$	$k_{44} = 7.7 \times 10^{-9} \exp\left(\frac{-17560}{T}\right)$	Sidhu *et al.* (1992)

Table 2.5: Analytical fits to the rate coefficients for the reactions determining the H_3^+ abundance.

Other simplifications

Making these simplifications leaves us with only a small number of chemicals to deal with. Nevertheless, the number of possible reactions is still large. Many of these reactions, however, have little effect on the final abundances: for instance, the small abundances of H^- and H_2^+ ensure that the mutual neutralization reaction

$$H^- + H_2^+ \rightarrow H_2 + H \tag{2.11}$$

will always be negligible (Abel *et al.*, 1997). The same is true for many of the other possible reactions. The task of identifying those reactions which can be safely neglected is long and involved but has already been performed to a great extent by Abel *et al.* (1997). In general, I agree with their decisions and make the same simplifications. However, a few points of disagreement should be noted.

Firstly, their assertion that secondary ionization can be neglected is true only in optically thin gas, where its effects are relatively small. In gas which is optically thick to ionizing radiation, however, secondary ionization significantly alters the ionization rate and must be included.

Secondly, they argue that collisional detachment of H^- by He or H_2 can be neglected because the reaction cross-sections are orders of magnitude lower than those for the corresponding reaction with hydrogen (reaction 15). However, this does not appear to actually be the case. The AMDIS Aladdin database [2] gives values for all three cross-sections that agree to within a factor of a few, while the rate coefficients given by Huq *et al.* (1982) and Huq *et al.* (1983), for He and H_2 respectively, are comparable to that for hydrogen. The reaction with H_2 can nevertheless be omitted due to the small H_2 abundance, but the reaction with He should be included; it becomes reaction number 29.

Thirdly, it is necessary to include the charge transfer reactions

$$He^+ + H \rightarrow He + H^+ \tag{2.12}$$

and

$$He + H^+ \rightarrow He^+ + H \tag{2.13}$$

in order to determine the He^+ abundance correctly. This is important because,

[2] http://www-amdis.iaea.org/aladdin.html

as demonstrated in section 2.3, He^+ recombination can be a significant secondary source of ionizing radiation.

It also proves necessary to distinguish between the two photodissociation reactions

$$H_2 + \gamma \rightarrow H_2^+ + e^- \tag{2.14}$$

(reaction 24a) and

$$H_2 + \gamma \rightarrow H + H^+ + e^- \tag{2.15}$$

(reaction 24b). In optically thin gas, the first will generally dominate, but if the gas is optically thick then both will be significant.

On the other hand, I omit the photodissociation of H_2 by photons above the Lyman limit (which Abel *et al.* include as reaction 28). In our applications it will always be negligible compared to photoionization (reaction 24) or two-step photodissociation (reaction 27); the reasons for this are discussed in section 2.4.

Finally, Abel *et al.* include the mutual neutralization reaction mentioned above but demonstrate that it will always be negligible. I omit it from my model.

Having made all of these simplifications, we are left with a model containing nine chemicals and thirty reactions. These reactions are discussed in more detail in the following sections.

2.3 Collisional chemistry

$H + e^- \rightarrow H^+ + 2e^-$
$He + e^- \rightarrow He^+ + 2e^-$
$He^+ + e^- \rightarrow He^{++} + 2e^-$
The rates for all three collisional ionization reactions are taken from Abel *et al.* (1997). The first two are based upon Janev *et al.* (1987), while the third is based on data from the AMDIS database. They have been calculated as functions of T_e, the gas temperature expressed in units of eV. These fits should be accurate to within a couple of percent. Less accurate fits, but with a much simpler functional form, are provided in Voronov (1997).

$H^+ + e^- \rightarrow H + \gamma$
The rate of recombination can be calculated exactly for hydrogen and for hydrogenic ions such as He^{++}. In practice, since every excited state contributes,

some degree of approximation is necessary. Nevertheless, highly accurate rates can be calculated. I adopt the fits by Hui & Gnedin (1997) to the calculations of Ferland *et al.* (1992). These should be accurate to within 2% for all temperatures of interest.

The recombination rates for both case A and case B recombination are listed. Case A recombination includes recombination to all bound states; case B omits recombinations directly to the ground state. Case A generally applies in gas which is optically thin to ionizing radiation. On the other hand, case B is appropriate in optically thick conditions, when we can assume that the ionizing photons produced by recombination directly to the ground state are absorbed immediately by other hydrogen atoms. This is a good approximation as long as the mean free path of the ionizing photons, given by

$$l = (\sigma_0 n_{\rm H})^{-1} \simeq 0.05 n_{\rm H}^{-1} \ {\rm pc}, \tag{2.16}$$

is small compared to any length scale of interest.

$\mathbf{He^+ + e^- \rightarrow He + \gamma}$

The recombination rate for He^+ cannot be calculated precisely, but can be approximated to high accuracy (Drake, 1996). The rates listed are fits to the tabulated data of Hummer & Storey (1998), who have calculated recombination rates for a range of temperatures between 10 K and 25000 K, with an accuracy comparable to that of the hydrogenic rates. At temperatures greater than 25000 K, the fits become extrapolations, but comparison with the rates listed in Hui & Gnedin (1997) (which remain valid until $T = 5 \times 10^5$ K) suggests that the accuracy remains reasonable. In any case, the high temperature recombination rate becomes dominated by dielectronic recombination. This involves the excitation of the bound electron during recombination and is not included in the calculations of Hummer & Storey. I list its contribution as a separate rate, k_{5di}, which is taken from Aldrovandi & Pequignot (1973) and which should be accurate to within 5%.

The treatment of He^+ recombination in optically thick gas involves an added complication. The difference in energy between the least energetic excited state of He^+ and the ground state is greater than 13.6 eV. Consequently, He^+ recombination produces photons capable of ionizing hydrogen. A detailed discussion of this effect is given in Osterbrock (1989); I summarize it here.

Recombination directly to the ground state produces photons which can ionize both hydrogen and helium. A fraction y of these photons will be absorbed by

hydrogen, where

$$y = \frac{n_{\rm H}\sigma_{20}(\nu_{\rm th})}{n_{\rm H}\sigma_{20}(\nu_{\rm th}) + n_{\rm He}\sigma_{21}(\nu_{\rm th})} \tag{2.17}$$

and $h\nu_{\rm th} = 24.6\,{\rm eV}$ is the ionization threshold of He I. For primordial gas with a small fractional ionization and low H_2 abundance, $n_{\rm He}/n_{\rm H} \simeq 0.08$, and $y = 0.677$.

The effects of recombination to the excited states are more involved, and depend upon the gas density. However, in the low density limit, 96% of all recombinations result in the production of a photon capable of ionizing hydrogen.

The effects of dielectronic recombination are similar. Excitation of the bound electron ensures that the photon produced during recombination is unlikely to be able to ionize hydrogen, but the subsequent decay of the excited electron will frequently produce an ionizing photon; for simplicity, we again take the probability to be 96%. It should be noted, however, that ionization of hydrogen due to dielectronic recombination is unlikely to ever be significant – generally, hydrogen will be collisionally ionized at the temperatures at which dielectronic recombination dominates.

Combining these results, we see that the He^+ recombination rate in optically thin gas is given by

$$k_{\rm tot} = k_{5a} + k_{5di} \tag{2.18}$$

whereas in optically thick gas it is

$$k_{\rm tot} = y(k_{5a} - k_{5b}) + k_{5b} + k_{5di} \tag{2.19}$$

In the latter case, it is accompanied by ionization of hydrogen at a rate

$$R_{\rm ion} = [y(k_{5a} - k_{5b}) + 0.96(k_{5b} + k_{5di})]\,\frac{n_{\rm e}n_{\rm He^+}}{n_{\rm H}}\ {\rm s}^{-1} \tag{2.20}$$

The importance of this contribution to the ionization rate will depend upon the ionization history of the gas. However, we can gain some insight into its likely importance by making a couple of simple approximations. If we assume that the gas is in photoionization equilibrium, and that charge transfer (and other chemical processes) can be neglected, then

$$R_{\rm He}n_{\rm He} = k_{\rm tot}n_{\rm e}n_{\rm He^+}, \tag{2.21}$$

where $R_{\rm He}$ is the He I ionization rate. Using this relation, we can write the ratio of $R_{\rm ion}$ to $R_{\rm H}$ (the ionization rate of hydrogen due to unprocessed radiation) as

$$\frac{R_{\rm ion}}{R_{\rm H}} = \frac{y(k_{5a} - k_{5b}) + 0.96(k_{5b} + k_{5di})}{k_{\rm tot}}\frac{R_{\rm He}}{R_{\rm H}}\frac{n_{\rm He}}{n_{\rm H}}. \tag{2.22}$$

Evaluating this, we find that

$$\frac{R_{\rm ion}}{R_{\rm H}} \sim 0.1 \frac{R_{\rm He}}{R_{\rm H}} \tag{2.23}$$

for all temperatures of interest. In optically thick gas illuminated by a hard ionizing spectrum, $R_{\rm He}$ will often be greater than $R_{\rm H}$ and thus He^+ recombination can be an important source of ionization. [3]. Including the effects of charge transfer (reactions 30 and 31) will reduce this effect somewhat.

$He^{++} + e^- \rightarrow He^+ + \gamma$

As He^{++} is a hydrogenic ion, its recombination rates can be obtained by a simple rescaling of those for H^+. As before, case A is appropriate for optically thin conditions, while in optically thick gas we must consider the fate of the emitted photons.

Photons produced by recombination to the ground state can ionize H, He or He^+ and their fate will depend upon the relative abundances of these three species. If the fractional ionization is small, then $n_{\rm He^+} \ll n_{\rm He}$ and all of the photons produced are absorbed by neutral hydrogen and helium. Consequently, recombination of He^+ proceeds at the case A rate even in optically thick gas. However, as the abundance of He^{++} will be even smaller than that of He^+, the photons produced by its recombination are unlikely to contribute significantly to the ionization rate.

$H + e^- \rightarrow H^- + \gamma$

A fit from Abel *et al.* (1997), based upon the photodetachment cross-sections of Wishart (1979) and de Jong (1972), from which it can be obtained via the principle of detailed balance. The resulting rate should be accurate to within 1% for $100 < T < 10^4$ K. At higher temperatures, the accuracy degrades, but this is unimportant; the rate is large and H^- is rapidly destroyed.

$H^- + H \rightarrow H_2 + e^-$

The fit is from Galli & Palla to the recent calculations of Launay *et al.* (1991). Using the earlier results of Browne & Dalgarno (1969), Abel *et al.* find a rate which lies within 10% at low temperatures, but differs by as much as 50% at high temperatures. Note, however, that significant uncertainty in this rate will

[3] As an alternative to this slightly awkward treatment, we could simply adopt case A recombination rates, and solve explicitly for the radiative transfer of the ionizing photons produced. However, the computational cost of doing this would be considerable, while adding little to the accuracy

generally have very little effect on the H_2 abundance, as this reaction will always be very much faster than the formation of H^-.

$H + H^+ \rightarrow H_2^+ + \gamma$

The fit is by Galli & Palla to the data of Ramaker & Peek (1976). Recent calculations by Stancil *et al.* (1993) agree to within 3% over the temperature range.

$H_2^+ + H \rightarrow H_2 + H^+$

A rate independent of temperature and consistent with the experimental measurement of Karpas *et al.* (1979) is generally adopted. The uncertainty is *at least* 20% (corresponding to the error in this measurement). However, rates for this reaction differing by as much as a factor of six have been used in the literature (e.g. Culhane & McCray, 1995). Nevertheless, the speed of the reaction is such that any error in the final H_2 abundance is likely to be small.

$H_2 + H^+ \rightarrow H_2^+ + H$

A rate from Galli & Palla, based on the cross-section of Holliday *et al.* (1971); no estimate of the accuracy is given. Note that the corresponding rate in Abel *et al.* is incorrect at temperatures below 10^4 K; the cross-sectional data on which their fit is based is accurate only at higher temperatures. In practice, this is only significant if the fractional ionization is large, as the usual fate of the H_2^+ produced in this reaction is to reform H_2.

$H_2 + e^- \rightarrow 2H + e^-$

The rate is a fit by Stibbe & Tennyson (1999) to their own calculational results. No estimate of the accuracy is given. Their parameterization is inaccurate at temperatures below $T = 7200$ K; however, at lower temperatures this reaction proceeds extremely slowly and will not be important. Note that the rate of this reaction is strongly dependant on the vibrational state of the molecule; the LTE rate is two of orders of magnitude larger than the $v = 0$ rate.

$H_2 + H \rightarrow 3H$

This rate is the low density limit of the rate computed by Martin *et al.* (1996) using quasi-classical trajectory calculations. It is comprised of two contributions: one from excitation to a classically unbound state ($k_{\rm cd}$) and the other from the formation of quasi-bound states, which subsequently undergo dissociative tunneling ($k_{\rm dt}$). These rates have rms errors of 3.9% and 8.7% respectively; the total error is less than 10% and will decrease as classical dissociation comes to domi-

nate at high temperatures. While a fully quantal treatment would be desirable, it seems likely that it would make little difference to the rates – the neglect of purely quantal effects introduces significant inaccuracy into the quasi-classical treatment only for $T \lesssim 600$ K, at which temperature the dissociation rate is negligible.

$H^- + e^- \rightarrow H + 2e^-$

A fit by Abel *et al.* to data from Janev *et al.* (1987). This reaction becomes important only for $T > 10^4$ K; at lower temperatures, mutual neutralization (reaction 16) dominates.

$H^- + H \rightarrow 2H + e^-$

Also from Abel *et al.* and based on Janev *et al.* (1987). The rate at low temperature ($T \lesssim 10^3$ K) is based on extrapolated data and is therefore quite uncertain; however, at these temperatures the formation of H_2 (reaction 8) is considerably more likely to occur.

$H^- + H^+ \rightarrow 2H$

The rate of this reaction is uncertain by an order of magnitude, particularly at low temperatures. I use the rate from Galli & Palla, which is based on the cross-sectional data of Moseley *et al.* (1970), as fit by Peterson *et al.* (1971). Abel *et al.* and Shapiro & Kang (1987) base their rates on different experimental data and arrive at different results. A better determination of this rate is clearly desirable, particularly as it will be the main competitor with reaction 8 at low temperatures. Nevertheless, as demonstrated by Palla & Zinnecker (1987), this reaction significantly affects the H_2 abundance only when the fractional ionization is greater than 10^{-2}.

$H^- + H^+ \rightarrow H_2^+ + e^-$

A fit by Galli & Palla, derived from the cross-section measured by Poulaert *et al.* (1978). Abel *et al.* give a fit to the same data which differs by a few percent, well within the accuracy of the experimental data. In any event, this reaction will usually be unimportant as it is significantly less likely to occur than mutual neutralization (reaction 16).

$H_2^+ + e^- \rightarrow 2H$

A fit by Abel *et al.* to the results of calculations by Schneider *et al.* (1994). It is incorrect for $T < 100$ K, but the free electron fraction at these temperatures will be negligible. As with mutual neutralization (reaction 16), it significantly affects the H_2 abundance only for fractional ionizations greater than 10^{-2}.

$H^- + He \rightarrow H + He + e^-$

The rate is taken from Huq *et al.* (1982). The data available in the AMDIS database has quoted errors of between 10 and 25%; the error in this rate coefficient should be of the same order of magnitude.

$He^+ + H \rightarrow He + H^+ + \gamma$

$He + H^+ \rightarrow He^+ + H$

Charge transfer from He^+ to H must be included if we are to calculate the He^+ abundance (and hence the ionization rate due to He^+ recombination) correctly. The rate becomes comparable with that for He^+ recombination when the fractional ionization is small. The rate for reaction 30 is a fit by Stancil *et al.* (1998) to the data of Zygelman *et al.* (1989) for radiative charge transfer; the photons produced have energies of the order of 11 eV, the difference in the hydrogen and helium ionization potentials. In energetic collisions, non-radiative charge transfer, which leaves the helium atom in an excited state, is also possible. However, this has an activation energy of several eV and thus at low temperatures radiative charge transfer dominates.

At high temperatures, the reverse reaction also becomes important. Above 10^4 K, the adopted rate is a fit from Galli & Palla to the calculations of Kimura *et al.* (1993). At lower temperatures, however, this fit significantly overestimates the rate. I have therefore fitted the low temperature portion of the Kimura *et al.* data with an alternative formula which incorporates the correct exponential temperature dependance. This will inevitably be inaccurate, as it is based on only a few data points, but as the rate rapidly becomes negligible, the error in the resulting abundances should be small.

2.4 Photochemistry

Given the specific intensity $I_\nu(\theta, \phi)$ of the incident radiation field, we can obtain rate coefficients from the radiative cross-sections listed in table 2.1 by integrating over frequency and angle

$$k_i = \int_0^{2\pi} \int_0^{\pi} \int_0^{\infty} \frac{I_\nu \sigma_{\nu,i}}{h\nu} e^{-\tau_\nu} \sin\theta \, \mathrm{d}\nu \, \mathrm{d}\theta \, \mathrm{d}\phi. \tag{2.24}$$

For an isotropic radiation field, this simplifies to

$$k_i = 4\pi \int_0^{\infty} \frac{I_\nu \sigma_{\nu,i}}{h\nu} e^{-\tau_\nu} \, \mathrm{d}\nu. \tag{2.25}$$

Most of the photochemical reactions involve continuum absorption, and the corresponding integral is straightforward to calculate numerically. A significant exception is the two-step photodissociation of H_2 (reaction 27), also known as the Solomon process. This involves the excitation of the H_2 molecule to an excited electronic state, followed by its decay to the vibrational continuum of the ground state and consequent dissociation. It is not a continuum process; rather, it proceeds through absorption in a large number of discrete spectral lines. This complicates determination of the rate coefficient, as discussed below.

$\mathbf{H} + \gamma \rightarrow \mathbf{H^+} + \mathbf{e^-}$
$\mathbf{He} + \gamma \rightarrow \mathbf{He^+} + \mathbf{e^-}$
$\mathbf{He^+} + \gamma \rightarrow \mathbf{He^{++}} + \mathbf{e^-}$

The parameter ϵ in the hydrogenic cross-sections is given by

$$\epsilon = \sqrt{\frac{\nu}{\nu_{\rm th}} - 1}, \tag{2.26}$$

where the energy thresholds $h\nu_{\rm th}$ are listed in table 2.3. These cross-sections are exact in the non-relativistic limit.

For He, no exact cross-section is available. Instead, I use the analytical fit to the cross-section presented by Yan *et al.* (1998), which has been specifically constructed to be accurate at high energies as well as near the ionization threshold. They estimate that it should be accurate to within 2% for photon energies less than 600 eV, degrading to around 5% at higher energies.

At large photon energies, the double ionization of helium is possible:

$$\mathrm{He} + \gamma \;\rightarrow\; \mathrm{He^{++}} + 2\mathrm{e^-}. \tag{2.27}$$

This has a threshold energy of 79.0 eV, and its cross-section is typically a few percent of that for single ionization (for the same photon energy). As we show below, this is smaller than the uncertainty in the ionization rate arising from secondary ionization and we can safely ignore it.

For very large photon energies, Compton ionization becomes important. In this process, an energetic photon is inelastically scattered by a bound electron, which acquires enough energy to escape. As the Compton cross-section only becomes significant for photon energies greater than a few keV, a single photon can be responsible for many ionizations. In practice, however, the gas will almost always

be optically thin at these energies, and Compton ionization is therefore only important in the limit of very high column density.

Another, more important, effect of high energy photons is secondary ionization. This can result whenever the photo-electron produced during photoionization has enough kinetic energy to collisionally ionize other atoms or molecules. As the kinetic energy is given by

$$E = h\nu - h\nu_{\rm th}, \tag{2.28}$$

this can be significant for radiation with $h\nu \gg h\nu_{\rm th}$.

Secondary ionization calculations are complicated by the large number of possible destinations for the kinetic energy. As well as being lost to ionizations, it is also lost in excitations of hydrogen and helium (and H_2, if present). Additionally, a fraction is lost as heat.

The process has recently been studied by Dalgarno *et al.* (1999), who determine the fate of the energy for a wide variety of initial electron energies, gas compositions and fractional ionizations. Their ionization results are expressed in terms of the mean energy per ion pair W, defined as

$$W = \frac{E}{N_{\rm ion}} \tag{2.29}$$

where $N_{\rm ion}$ is the number of secondary ionizations produced. Clearly, given their determination of $W(E)$ for a particular type of ionization, we can calculate $N_{\rm ion}$ as a function of energy, and hence obtain the ionization rate

$$k_i = 4\pi \int_0^\infty \frac{I_\nu \sigma_{\nu,i}}{h\nu}(1 + N_{\rm ion})e^{-\tau_\nu}\,{\rm d}\nu. \tag{2.30}$$

The importance of secondary ionization depends upon the optical depth of the gas and the shape of the incident spectrum. The ionization cross-sections for hydrogen, helium and H_2 all fall off sharply with increasing frequency, so when the gas is optically thin, the ionization rate is dominated by photons near the ionization threshold. In this case, the effect of secondary ionization is small. In optically thick gas, however, the flux at these wavelengths is highly attenuated, the ionization rate is dominated by higher energy photons and the importance of secondary ionization is correspondingly greater. Thus, as the optical depth increases, the importance of secondary ionization grows.

Various sources of error are present in our determination of $N_{\rm ion}$, and hence in the ionization rate. To begin with, Dalgarno *et al.* (1999) present results for a limited

number of photon energies; values for other energies must be interpolated. The inaccuracy introduced by this interpolation is greatest at low energy, where W varies rapidly; at higher energies, W tends to a constant value and the inaccuracy is small. Other errors result from the fact that none of the hydrogen-helium mixtures considered by Dalgarno *et al.* is a perfect match to primordial gas – the closest match has a fractional helium abundance of 0.1, which is rather too high – and because all of the rates depend upon the fractional ionization, necessitating a further interpolation. Finally, a degree of uncertainty is inevitably present in the atomic data upon which the calculations are based. Taken together, these factors lead to an uncertainty of around 10 – 15% in the value of $N_{\rm ion}$. The resulting uncertainty in the ionization rate clearly depends upon the shape of the incident spectrum.

A final source of photoionization is the recombination of He^+ and He^{++}. This has already been discussed and I do not consider it further here.

$H^- + \gamma \rightarrow H + e^-$

A fitting formula from Shapiro & Kang (1987), accurate to within 10%. For frequencies between 2×10^{14} Hz and 2×10^{15} Hz the tabulated data of Wishart (1979) is also available, with an accuracy of better than 1%. The threshold energy for this reaction is 0.755 eV.

$H_2 + \gamma \rightarrow H_2^+ + e^-$

$H_2 + \gamma \rightarrow H + H^+ + e^-$

The threshold for photoionization of H_2 to H_2^+ is 15.42 eV, while above 18.08 eV, dissociative ionization also becomes possible. Frequently these two processes are combined, together with double photoionization,

$$H_2 + \gamma \rightarrow 2H^+ + 2e^- \tag{2.31}$$

into a single cross-section. However, this is not suitable for my purposes, as H_2 ionized to H_2^+ will quickly re-form via reaction 10, while that ionized to $H + H^+$ will not.

Accordingly, I adopt cross-sections of the form

$$\sigma_{24a} = [1 - f(x)]\ \sigma_{24} \tag{2.32}$$

for simple ionization and

$$\sigma_{24b} = f(x)\ \sigma_{24} \tag{2.33}$$

for dissociative ionization, where σ_{24} is the total H_2 photodissociation cross-section. This is listed in table 2.1, and is based on the work of O'Neil & Reinhardt (1978), Wilms *et al.* (2000) and Yan *et al.* (1998). The fraction of the total cross-section contributed by dissociative ionization, $f(x)$, can be calculated from the results of Samson & Haddad (1994); I find that

$$f(x) = \begin{cases} 0 & x < 0\,\mathrm{eV} \\ 4.34 \times 10^{-3} x & 0 < x < 50\,\mathrm{eV} \\ 0.22(x/50)^{-0.1} & x > 50\,\mathrm{eV} \end{cases} \tag{2.34}$$

where $x = E - E_0$, and $E_0 = 18.08\,\mathrm{eV}$ is the dissociative ionization threshold. This fit does not properly reproduce the structure in the cross-section at low energies ($E < 50\,\mathrm{eV}$) but in optically thick conditions, the contribution from this portion will be small.

This treatment neglects the contribution from double photoionization, but this is never larger than a few percent, comparable to the error in σ_{24}, and can be safely neglected.

$H_2^+ + \gamma \rightarrow H + H^+$

The fit given by Abel *et al.* to the calculations of Stancil (1994) appears to be erroneous; it gives a cross-section that is many orders of magnitude smaller than the true one. In its place, I use fits from Shapiro & Kang (1987), based on the data of Dunn (1968). These will be inaccurate at high energies, but this is unlikely to significantly affect the dissociation rate, which will be dominated by radiation below the Lyman limit.

This rate is appropriate when all of the H_2^+ is in the ground state, and will underestimate the dissociation rate at high redshift ($z \gtrsim 200$) where the CMB populates excited vibrational states.

$H_2^+ + \gamma \rightarrow 2H^+ + e^-$

Above a threshold of 30.0 eV, this alternative form of H_2^+ photodissociation is possible. The rate is taken from Shapiro & Kang (1987) and is based on the data of Bates & Öpik (1968).

$H_2 + \gamma \rightarrow H_2^* \rightarrow 2H$

The binding energy of H_2 is 4.476 eV, but photons of this energy have little chance of dissociating H_2 as the simplest dissociative transition – excitation to the vibrational continuum of the ground state – is strongly forbidden (Field *et al.*, 1966). Transitions to $b^3\Sigma_u^+$, the least energetic of the excited electronic states, are

also forbidden. Consequently, H_2 photodissociation can take place only through higher energy states.

The primary route for H_2 photodissociation is the Solomon process (Stecher & Williams, 1967). [4] This involves excitation to the Lyman ($B\,^1\Sigma_u^+$) or Werner ($C\,^1\Pi_u$) states followed (some fraction of the time) by radiative decay to the vibrational continuum of the ground state. As a number of vibrational levels are accessible in each excited state, photodissociation takes place through a number of discrete absorption lines, known as the Lyman and Werner band systems.

The Lyman and Werner states both contain a large number of bound rotational and vibrational levels. However, transitions to these levels are governed by selection rules which significantly reduce the number which are accessible. Specifically, transitions involving a change of parity must leave the rotational quantum number, J, unchanged; conversely, transitions involving no change of parity must change J by plus or minus one. There is no corresponding selection rule for the vibrational quantum number, although transitions requiring energies greater than 13.6 eV will be strongly suppressed by hydrogen continuum absorption in optically thick gas.

Applying these rules to the Lyman state, which has the same parity as the ground state, we find that:

- If the initial rotational quantum number $J_i \geq 1$ then only two rotational levels are accessible for each vibrational level in B.
- If $J_i = 0$, then only one rotational state ($J = 1$) is accessible for each vibrational level.

The Werner state is more complicated, as it is actually split into two states of very similar energy but differing parity, C^+ and C^-; C^+ has the same parity as the ground state. Additionally, there is no $J = 0$ level in C – the contribution of the electron orbital angular momentum ensures that the total angular momentum must be non-zero. In this case:

- If $J_i \geq 2$ then two rotational levels are accessible for each vibrational level in C^+, but only one per level in C^-.

[4] Named for P. M. Solomon, who first drew attention to its importance.

- If $J_i = 1$ then there is only one accessible rotational level per vibrational level in both C^+ *and* C^-.

- If $J_i = 0$ then there is again only one accessible rotational level per vibrational level in C^+, while C^- is inaccessible.

A little consideration shows that molecules which are excited to the Lyman or Werner states and then decay to the ground state preserve their character as either ortho or para-hydrogen.

The total photodissociation rate can be written as

$$k_{\rm dis} n_{\rm H_2} \equiv \sum_l k_{{\rm dis},l} n_l \tag{2.35}$$

where n_l is the number density of molecules in some rovibrational level l of the ground state, and $k_{{\rm dis},l}$ is the photodissociation rate due to transitions out of l. The latter is given in turn by

$$k_{{\rm dis},l} = \sum_u \zeta_{ul} f_{{\rm dis},u} \tag{2.36}$$

where ζ_{ul} is the pumping rate – the rate of transitions from l to some level u in one of the excited states – and $f_{{\rm dis},u}$ is the dissociation probability – the fraction of decays from u which end in the vibrational continuum of the ground state and hence result in dissociation. The pumping rate for a particular transition is given by

$$\zeta_{ul} = \int_0^{2\pi} \int_0^{\pi} \int_0^{\infty} \frac{I_\nu \sigma_\nu}{h\nu} e^{-\tau_\nu} \sin\theta \, {\rm d}\nu \, {\rm d}\theta \, {\rm d}\phi \tag{2.37}$$

where σ_ν is the radiative cross-section for that transition. If we ignore the very small correction for stimulated emission, then σ_ν can be written as

$$\sigma_\nu = 2.654 \times 10^{-2} f_{\rm osc} \phi_\nu \ {\rm cm}^{-2}, \tag{2.38}$$

where $f_{\rm osc}$ is the oscillator strength and ϕ_ν the line profile.

Optically thin gas

At frequencies below the Lyman limit, the continuum opacity of primordial gas is very small (Lenzuni *et al.*, 1991) and τ_ν is determined almost entirely by H_2

self-absorption. [5] When the H_2 column density is small, this is negligible and we can write the pumping rate as

$$\zeta_{ul} = \int_0^{2\pi} \int_0^{\pi} \int_0^{\infty} \frac{I_\nu \sigma_\nu}{h\nu} \sin\theta \, \mathrm{d}\nu \, \mathrm{d}\theta \, \mathrm{d}\phi \tag{2.39}$$

If we assume that I_ν is isotropic, and that the line widths are sufficiently small that the line profile can be approximated as a delta function, then

$$\zeta_{ul} = 0.334 f_{\mathrm{osc}} \frac{I(\nu_{ul})}{h\nu_{ul}}, \tag{2.40}$$

where ν_{ul} is the frequency at line centre.

With these approximations, the photodissociation rate out of level l becomes

$$k_{\mathrm{dis},l} = \sum_u 0.334 f_{\mathrm{osc},u} f_{\mathrm{dis},u} \frac{I(\nu_{ul})}{h\nu_{ul}}. \tag{2.41}$$

This rate depends only on the incident spectrum and atomic data, and thus is independent of the density and temperature of the gas. The total rate, on the other hand, depends upon the populations of the various excited levels of H_2, which themselves depend upon density and temperature. We know from the discussion in section 2.2 that the populations of the excited vibrational levels are always very small; the same is not true of the excited rotational levels.

Nevertheless, let us assume for the moment that all of the H_2 is in the form of para-hydrogen, with $J = 0$. In this case, the total photodissociation rate is simply

$$k_{\mathrm{dis}} = 0.334 \sum_{u=1}^{25} f_{\mathrm{osc},u} \, f_{\mathrm{dis},u} \frac{I(\nu_{ul})}{h\nu_{ul}} \tag{2.42}$$

where we sum over the 25 possible transitions which have energies less than 13.6 eV.

The value of k_{dis} clearly depends upon the incident spectrum. If I_ν is constant, then it can be taken out of the sum and calculation of k_{dis} becomes trivial. Using oscillator strengths and frequencies from Abgrall *et al.* (1993a,b) and Roueff (1997), plus dissociation probabilities from Abgrall *et al.* (2000), we find that

$$k_{\mathrm{dis}} = 1.38 \times 10^9 \, I_\nu \, \mathrm{s}^{-1}. \tag{2.43}$$

[5] Absorption in the Lyman-series lines of atomic hydrogen also becomes important at high column densities.

This is in good agreement with the value obtained by Abel *et al.* (1997) under similar assumptions. [6]

This expression remains a good approximation in the more general case where I_ν is allowed to vary over the range of wavelengths encompassing the Lyman-Werner bands, as long as we replace I_ν with $I(\bar{\nu})$, where $h\bar{\nu} = 12.87\text{eV}$ (Abel *et al.*, 1997). This frequency corresponds to transitions to the $v = 13$ vibrational level of the Lyman state, which lies in the middle of the set of transitions that dominate the photodissociation rate. The accuracy of this approximation is surprisingly good, with the constant of proportionality varying by no more than a few percent for a wide range of different spectra.

If we relax the assumption that all of the H_2 has $J = 0$, little changes. More transitions become available, but each set of transitions has a smaller number of absorbing molecules associated with it, so the overall change in the photodissociation rate is small. For gas with rotational levels in LTE at a temperature of 1000 K, the photodissociation rate rises by less than 10%, to

$$k_{\text{dis}} = 1.50 \times 10^9 \, I(\bar{\nu}) \, \text{s}^{-1}. \tag{2.44}$$

Optically thick gas

So far, we have assumed that the gas is optically thin to Lyman-Werner band radiation. For H_2 column densities $N_{H_2} \gtrsim 10^{14} \, \text{cm}^{-2}$, however, the stronger lines begin to self-shield and this assumption breaks down. My treatment in this case follows that of Draine & Bertoldi (1996). We begin by writing the pumping rate as

$$\zeta_{ul} = \int_0^{4\pi} \int_0^{\infty} \frac{I_\nu \sigma_\nu}{h\nu} e^{-\tau_\nu} \, \text{d}\nu \, \text{d}\Omega, \tag{2.45}$$

where $\text{d}\Omega = \sin\theta \, \text{d}\theta \, \text{d}\phi$ is an element of solid angle. If the radiation field is isotropic, then the angular dependence of this equation comes purely from the optical depth, which can be written as

$$\tau_\nu = \int_0^R \sigma_\nu(r) n_l \, \text{d}r, \tag{2.46}$$

for any particular line of sight. Here r is the distance along the line of sight, and R the distance to the boundary of the protogalaxy. In general, σ_ν depends on r

[6] Note that their parameter j_ν corresponds to $4\pi I_\nu$.

in two ways – through local variations of the temperature or small-scale velocity field, that change the Doppler width of the line, and through the effect of any large-scale velocity gradient, which can cause absorption from distant gas to be Doppler-shifted relative to that from nearby gas. If we assume, however, that both of these are negligible, then equation 2.46 reduces to

$$\tau_\nu = \sigma_\nu N_l, \tag{2.47}$$

where N_l is the column density of molecules in level l along the line of sight. In this case, the pumping rate becomes

$$\zeta_{ul} = \int_0^{4\pi} \int_0^{\infty} \frac{I_\nu \sigma_\nu}{h\nu} e^{-\sigma_\nu N_l} \, \mathrm{d}\nu \, \mathrm{d}\Omega. \tag{2.48}$$

Alternatively, we can write this as

$$\zeta_{ul} = \int_0^{4\pi} \frac{\mathrm{d}\zeta_{ul}}{\mathrm{d}\Omega} \, \mathrm{d}\Omega, \tag{2.49}$$

where

$$\frac{\mathrm{d}\zeta_{ul}}{\mathrm{d}\Omega} = \int_0^{\infty} \frac{I_\nu \sigma_\nu}{h\nu} e^{-\sigma_\nu N_l(\Omega)} \, \mathrm{d}\nu. \tag{2.50}$$

As long as I_ν does not vary significantly over the width of the line, this can be simplified to

$$\frac{\mathrm{d}\zeta_{ul}}{\mathrm{d}\Omega} = \frac{I_{\nu_0}}{h} \int_0^{\infty} \sigma_\nu e^{-\sigma_\nu N_l(\Omega)} \frac{\mathrm{d}\nu}{\nu}. \tag{2.51}$$

If we now make the identification (Draine & Bertoldi, 1996)

$$\frac{\mathrm{d}W_{ul}}{\mathrm{d}N_l} = \int_0^{\infty} \sigma_\nu e^{-\sigma_\nu N_l} \frac{\mathrm{d}\nu}{\nu}, \tag{2.52}$$

where W_{ul} is the dimensionless equivalent width of the line

$$W_{ul}(N_l) = \int_0^{\infty} (1 - e^{-\sigma_\nu N_l}) \frac{\mathrm{d}\nu}{\nu}, \tag{2.53}$$

then the pumping rate becomes

$$\frac{\mathrm{d}\zeta_{ul}}{\mathrm{d}\Omega} = \frac{I(\nu_0)}{h} \frac{\mathrm{d}W_{ul}}{\mathrm{d}N_l}. \tag{2.54}$$

Summing over levels, we find that the total photodissociation rate becomes

$$k_{\mathrm{dis}} = \sum_u \sum_l \frac{I(\nu_{ul})}{h} f_{\mathrm{dis},u} \frac{n_l}{n_{\mathrm{H_2}}} \int_0^{4\pi} \frac{\mathrm{d}W_{ul}}{\mathrm{d}N_l} \, \mathrm{d}\Omega. \tag{2.55}$$

Thus, to determine the photodissociation rate, we need to know the level populations n_l and the associated column densities $N_l(\Omega)$; the latter can be used to calculate $\mathrm{d}W_{ul}/\,\mathrm{d}N_l$, by means of the numerical approximations presented in Rodgers & Williams (1974).

In the limit of large column densities, this prescription must be adjusted, to prevent W_{ul} increasing without limit. As noted by Draine & Bertoldi (1996), this unphysical behaviour is due to our neglect of line overlap (and of absorption by neutral hydrogen). Following Draine & Bertoldi, we can account for these effects by making use of a modified equivalent width, $\tilde{W}$, defined by

$$\frac{\mathrm{d}\tilde{W}_{ul}}{\mathrm{d}N_l} = \frac{\mathrm{d}W_{ul}}{\mathrm{d}N_l} \exp(-W_{\mathrm{tot}}/W_{\mathrm{max}}), \tag{2.56}$$

where

$$W_{\mathrm{tot}} = \sum_l \sum_u W_{ul} + \sum_{n=3}^{15} W_{\mathrm{H}}(1 \rightarrow n). \tag{2.57}$$

and W_{max} is the total equivalent width of the portion of the spectrum containing the Lyman-Werner bands – this is roughly $1110 > \lambda > 912\,\text{Å}$, so $W_{\mathrm{max}} = \ln(1110/912) \simeq 0.2$. Our expression for W_{tot} includes the equivalent width of all of the Lyman lines from Lyβ up to $n = 15$; the higher lines blend together into a continuum. These widths are calculated using transition probabilities from Brocklehurst (1971).

Since the photodissociation rate depends upon the level populations, this means that, in theory, we should calculate these for all of the bound levels of H_2. In practice, this is unnecessary; as we saw previously, the only levels to acquire significant populations are the first few rotational levels. Nevertheless, the need to track even these levels still imposes a significant burden of extra work – as we shall see later, the small equilibrium timescales associated with the rotational levels force us to use very short timesteps to simulate the evolution of the gas. It is therefore useful to investigate the contribution of these levels to the photodissociation rate – how much difference do they actually make?

In figure 2.4, I plot the photodissociation rate obtained for 1000 K gas for four different assumptions about the level populations. All four rates are normalized by their optically thin values, which differ by only a few percent. The solid line assumes that all of the H_2 is in the form of para-H_2 and in the ground state; this corresponds to the assumption made by Abel *et al.* (1997). The dashed line assumes that the ortho to para ratio is 3:1, but again assumes that all of the H_2 is

in the ground state. The dot-dashed line assumes that the level populations are those set by statistical equilibrium between collisional excitation and collisional and radiative de-excitation in gas with neutral hydrogen number density $n_{\rm H} = 1\ {\rm cm}^{-3}$, which is a reasonable approximation for protogalactic gas outwith the core regions. Finally, the dotted line assumes that the level populations have their thermal equilibrium populations.

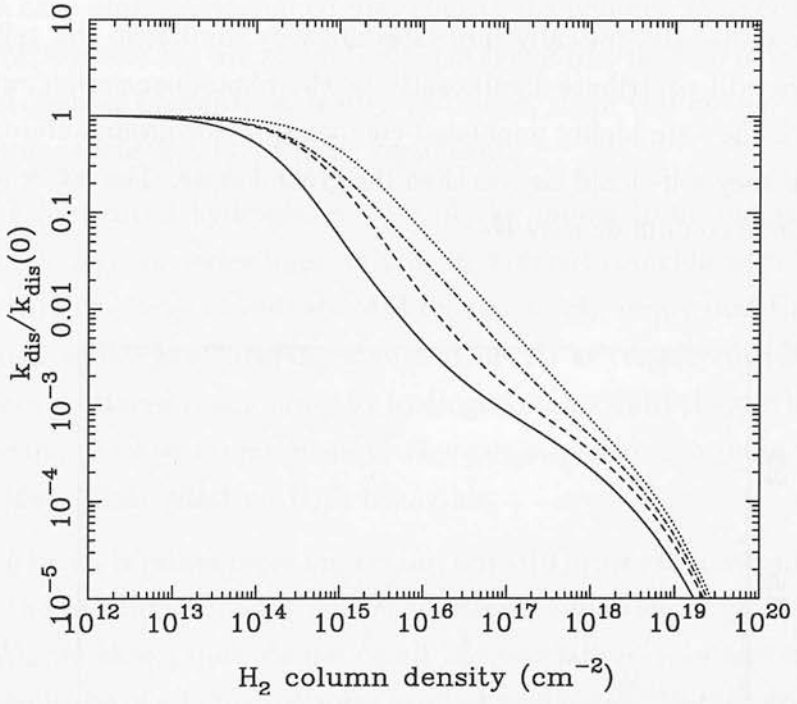

Figure 2.4: Photodissociation rates for H_2 for various different sets of level populations – pure para-hydrogen with no excited levels (solid line); a 3:1 ortho-para ratio, but no other excited levels (dashed line); statistical equilibrium at 1000 K, for a neutral hydrogen number density $n_{\rm H} = 1\ {\rm cm}^{-3}$ (dot-dashed line); thermal equilibrium at 1000 K (dotted line). In each case, an H_2 abundance of 10^{-4} is assumed.

At low column densities, we already know that the level populations make little difference. Figure 2.4 demonstrates that the same is also true at high column densities, where absorption by neutral hydrogen causes the photodissociation rate to fall off sharply. [7] For intermediate column densities ($10^{14} < N_{\rm H_2} < 10^{18}\,{\rm cm}^{-2}$), the differences are much more pronounced. The reason for this is easily explained.

[7]Note that my calculations assume an H_2 abundance of 10^{-4}, and thus the H_2 column density at which the rates fall off sharply, $N_{\rm H_2} = 10^{19}\,{\rm cm}^{-2}$, corresponds to a neutral hydrogen column density of $N_{\rm H} = 10^{23}\,{\rm cm}^{-2}$.

We can write the contribution of any particular level to the photodissociation rate as

$$k_i = \zeta_{i,0} f_i n_i \tag{2.58}$$

where $\zeta_{i,0}$ is the dissociation rate in optically thin gas, and f_i is a correction for self-shielding. This can be written in terms of the photodissociation rate from the ground state as

$$\frac{k_i}{k_0} = \frac{\zeta_{i,0}}{\zeta_{0,0}} \frac{f_i}{f_0} \frac{n_i}{n_0}. \tag{2.59}$$

Now, we know that the optically thin rates are very similar, so this tells us that excited states will contribute significantly to the photodissociation rate in two situations – if they are highly populated compared to the ground state, which is unlikely, or if they self-shield far less than the ground state. The latter is the case for intermediate column density H_2.

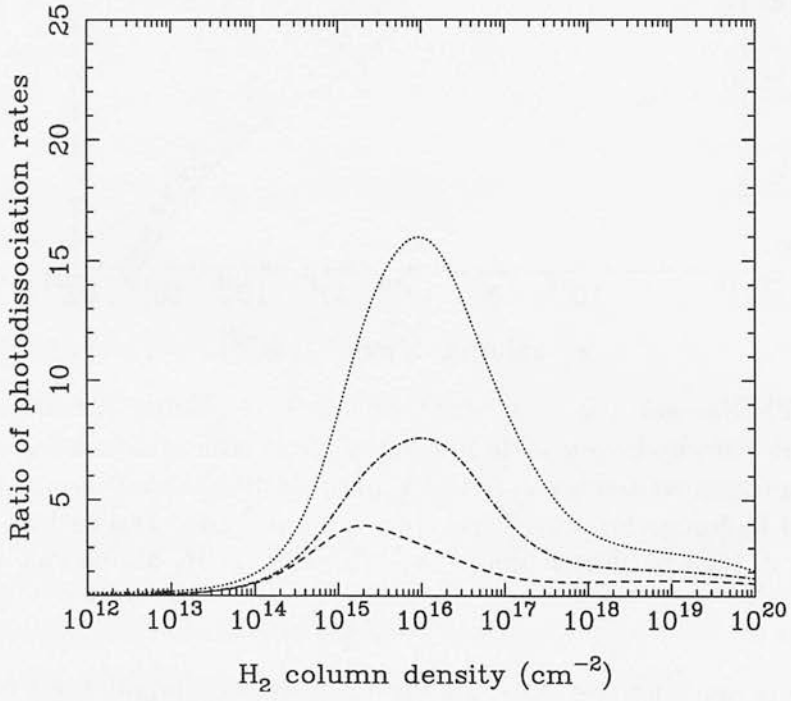

Figure 2.5: The ratio of the photodissociation rates plotted in figure 2.4.

Another view of the magnitude of this effect is given in figure 2.5, where I have plotted the various photodissociation rates as ratios of the smallest one (*i.e.* that for pure para-hydrogen, with no excitation). We see that in the worst case, the difference in the rates is more than an order of magnitude. In reality, of course,

we would generally model the H_2 as a mixture of ortho and para-hydrogen, and the true level populations would more closely resemble the statistical equilibrium ones, so the discrepancy would not be as large. I shall return to this point in chapter 4.

Scattering

So far we have implicitly assumed that none of the photons absorbed, whether by neutral hydrogen or H_2, are re-emitted; in other words, that there is no scattering of Lyman-Werner photons. In reality, of course, there will be *some* scattering – the question is whether this will be significant.

Scattering by neutral hydrogen can readily be shown to be unimportant. The fact that the Lyman series lines do not, in general, coincide with the H_2 lines, together with the large abundance of H relative to H_2, imply that the probability of a Lyman series photon being absorbed by H_2 is very small. Rather, these photons are scattered many times by hydrogen atoms until they either escape, or are converted to lower energy photons (Lyman-α, plus one or more visible/infrared photons). Their effect on H_2 is negligible.

Scattering by H_2 is rather more important, but still proves to have only a small net effect on the photodissociation rate. The reason is that although roughly 85% of Lyman-Werner absorptions do not result in dissociation, very few of these result in the re-emission of a Lyman-Werner band photon either. Rather, the majority of excited H_2 molecules decay first to an excited vibrational level in the ground state, producing a photon redwards of 1100 Å, only decaying thereafter to the ground state. Moreover, the probability of a re-emitted photon subsequently causing a dissociation is again only 15%. Thus the net contribution to the photodissociation rate is small, and can generally be ignored.

Photodissociation above the Lyman limit

So far, I have considered only photons below the Lyman limit. Including those above the Lyman limit would result in an increased photodissociation rate, as more vibrational and rotational levels in the Lyman and Werner states become accessible. Additionally, various other photodissociation reactions become possible at higher energies. These include two-step photodissociation via higher energy

Rydberg states such as $B'\,^1\Sigma_u^+$ (Abgrall *et al.*, 2000), or absorption into the continua of the Lyman and Werner states (Allison & Dalgarno, 1969; Glass-Maujean, 1986).

A complete treatment of H_2 photodissociation should include all of these effects. However, in practice, there is little to be gained by doing this. All of these processes are effective only near the Lyman limit, where absorption by neutral hydrogen will strongly attenuate the flux. At higher energies, H_2 photodissociation (reaction 24) dominates, by several orders of magnitude. Since, in the protogalaxies studied in this thesis, there is always far more atomic than molecular hydrogen, this suggests that photodissociation by any of these routes will never be important and can be neglected.

2.5 The cooling function

Although primordial gas chemistry is clearly of interest in its own right, our main concern is its effects upon the thermal evolution of protogalactic gas. Accordingly, in this section I briefly discuss the processes responsible for heating and cooling the gas. These are summarized in table 2.6. Briefly, cooling at low temperatures ($T < 10^4$ K) is dominated by molecular hydrogen, at intermediate temperatures (10^4 K $<$ T $<$ 10^6 K) by collisional excitation of hydrogen and helium, and at high temperatures ($T > 10^6$ K) by thermal bremsstrahlung. The other processes listed are generally less important.

2.5.1 Collisional excitation

The basis of collisional excitation cooling is simple: collisional excitation of an excited state, followed by radiative de-excitation, results in the production of one or more photons. As long as these photons can escape from the gas, the net result will be a loss of thermal energy.

In optically thin conditions, the cooling rate per unit volume due to transitions from an upper state u to a lower state l is given by

$$\Lambda_{ul} = n_u A_{ul} E_{ul}, \tag{2.60}$$

Collisional excitation		
H	$\Lambda_1 = 7.5 \times 10^{-19} \left(1 + T_5^{1/2}\right)^{-1} \exp\left(\frac{-118348}{T}\right) n_{\rm e}\, n_{\rm H}$	Cen (1992)
He	$\Lambda_2 = 1.85 \times 10^{-29} T^{0.082} \exp\left(\frac{-230000}{T}\right) n_{\rm e}\, n_{\rm He}$	Bray *et al.* (2000)
He^+	$\Lambda_3 = 5.6 \times 10^{-17} T^{-0.33} \exp\left(\frac{-473636}{T}\right) n_{\rm e}\, n_{\rm He^+}$	Aggarwal *et al.* (1992)
H_2	See text	—
Chemical changes		
	See text	—
Bremsstrahlung		
	$\Lambda_4 = 1.426 \times 10^{-27} Z_{\rm i}^2 \langle g_{ff} \rangle\, n_{\rm e}\, n_{\rm i}$	Spitzer (1978)
Compton scattering		
	$\Lambda_5 = 1.017 \times 10^{-37} T_{\rm r}^4 (T - T_{\rm r})\, n_{\rm e}$	Peebles (1993)
Photoionization/photodissociation		
	See text	

Table 2.6: The main sources of heating and cooling in primordial gas. All rates are in units of erg $\rm cm^{-3}\, s^{-1}$.

where n_u is the number density of particles in the upper state, A_{ul} is the spontaneous decay rate and E_{ul} is the energy of the transition. To link this to the collisional excitation rate, we need to understand how the upper state is populated.

If we assume the various states to be in statistical equilibrium, then at low densities we have

$$n_u A_{ul} = q_{lu} n_l n_c, \tag{2.61}$$

where q_{lu} is the collisional excitation rate for collisions with a partner of number density n_c, and we have assumed that radiative pumping is unimportant. In this case, the cooling rate becomes

$$\Lambda_{ul} = q_{lu} E_{ul} n_l n_c. \tag{2.62}$$

The total line cooling rate can then be obtained by summing over all possible transitions. [8]

At higher densities, collisional de-excitation cannot be neglected and the cooling rate must be modified. If we again assume statistical equilibrium, then we have

$$n_u A_{ul} + q_{ul} n_u n_c = q_{lu} n_l n_c \tag{2.63}$$

and the cooling rate becomes

$$\Lambda_{ul} = \frac{A_{ul} E_{ul} n_l q_{lu} n_c}{A_{ul} + q_{ul} n_c} \tag{2.64}$$

which can be written in terms of the low-density cooling rate, $\Lambda_{ul}(n \to 0)$, as

$$\Lambda_{ul} = \frac{\Lambda_{ul}(n \to 0)}{1 + n_c / n_{\rm crit}} \tag{2.65}$$

where $n_{\rm crit} = A_{ul}/q_{ul}$. Again, the total rate is given by summing over the rates for the individual transitions, recalling that the value of $n_{\rm crit}$ depends upon the transition.

If the optical depth of the emission lines is large, this can also significantly affect the cooling rate, as the emitted photons will be absorbed and re-emitted many times before escaping from the gas. As long as they escape *eventually*, this does not matter. However, in the period between each absorption and re-emission, there is a small chance that the absorber will undergo collisional de-excitation, returning the energy of the photon to the gas.

If we denote the probability that a photon will escape before being absorbed as p_e and the probability of collisional de-excitation as p_d, then, in the limit of a large number of scatterings, a fraction

$$f_{\rm esc} = \frac{p_e}{p_e + p_d} \tag{2.66}$$

of the photons will eventually escape from the gas. At any point in the gas, p_e is given by

$$p_e = \int_0^{4\pi} \int_0^{\infty} \phi_\nu e^{-\tau_\nu} \sin\theta \, {\rm d}\nu \, {\rm d}\Omega \tag{2.67}$$

where ϕ_ν is the line profile. In general, the value of this integral will depend upon a number of factors, such as the position within the cloud and the details of the

[8] Note that this assumes that only one type of collision is significant. This is frequently the case – for example, collisions with electrons dominate the atomic line cooling rates – but if not, it is straightforward to adapt the equations.

velocity field. However, a good approximation, valid when the emission lines are strong and there is no coherent large-scale velocity field, is given by (Emerson, 1996)

$$p_e = \frac{2}{3\tau_0}, \tag{2.68}$$

where τ_0 is the optical depth at line centre.

Compared to this, p_d is much simpler to calculate, being given at any point by

$$p_d = \frac{q_{\rm tot} n_c}{A_{\rm tot} + q_{\rm tot} n_c} \tag{2.69}$$

where $A_{\rm tot}$ and $q_{\rm tot}$ are the spontaneous and collisional transition rates, summed over all available states.

Hydrogen and helium

Radiative transitions in H, He and He^+ occur much faster than collisional transitions for all densities in our range of interest. Consequently, p_d is very small, and thus although p_e is also small, we find that $f_{\rm esc} \simeq 1$. Moreover, the large transition probabilities also ensure that only the ground state of each atom is significantly populated, thereby substantially simplifying the calculation of the cooling rates.

In cosmological applications, the most commonly used set of cooling rates is that first presented by Black (1981) and corrected for high temperatures by Cen (1992). However, the atomic data on which these rates are based is now at least two decades old, and it seems sensible to assess their accuracy.

In the case of hydrogen, the Black (1981) cooling rate is based upon direct calculation of the excitation rate to the $2s$ and $2p$ states, supplemented with approximate excitation rates for a large number of higher energy states. These are obtained by an appropriate scaling of the results for $2p$, as outlined in Gould & Thakur (1970).

Calculations of excitation rates to some of these excited states have undoubtedly improved over the past two decades, but current calculations still have uncertainties of between 10 and 20% and include only a handful of states (Callaway, 1994). Thus, while some improvement to the Black (1981) rate could be made, it is not clear that the gain in accuracy would be particularly great.

On the other hand, the cooling rates for He and He^+ are in need of revision. The values listed in Black are based upon excitation to a small number of states and thus underestimate the true cooling rate. In both cases, recent calculations allow for a substantially better treatment.

My cooling rate for He is derived from the work of Bray *et al.* (2000), which incorporates the first 28 excited states of He. The fit presented in table 2.6 is accurate to within 10% for temperatures in the range $3.75 \leq \log T \leq 5.75$; this is comparable to the accuracy of the underlying data. At low temperature, the cooling rate drops off exponentially and excitations from the small number of atoms in the metastable 2^3S level come to dominate. However, the effect of this is negligible compared to H excitation cooling at the same temperature, and need not be included.

For He^+, a cooling rate can be derived from the data of Aggarwal *et al.* (1992), valid for temperatures in the range $5000 < T < 5 \times 10^5$ K. This is based upon excitation to the first five excited states of He^+, and as such will still underestimate the true cooling rate, but it is a significant improvement over the rate given by Black (1981), which includes only the first *two* excited states. The fit may become inaccurate at higher temperatures, but at these temperatures most of the He^+ will have been collisionally ionized to He^{++}, and bremsstrahlung will dominate the cooling.

Molecular hydrogen

As dipole transitions between different rotational and vibrational levels of H_2 are forbidden, spontaneous transitions are quadrupolar, with correspondingly small transition probabilities. This has two consequences for H_2 cooling. Firstly, the optical depths of the rotational and vibrational lines are small, and line-trapping becomes significant only for very large H_2 column densities ($N_{H_2} \gtrsim 10^{23}$ cm^{-2}).

Secondly, the small transition probabilities allow significant populations to build up in excited levels at fairly moderate densities. Consequently, accurate treatment of H_2 cooling must include cooling from these levels as well as from the ground state.

Accurate determination of the H_2 cooling function is also hampered by the computational difficulties involved in accurate determinations of the collision rates.

In particular, the rate coefficients for H–H_2 collisions, which dominate the cooling, are highly sensitive to the detailed form of the interaction potential (Galli & Palla, 1998). In view of this, it is not surprising that a number of different calculations of the H_2 cooling rate exist in the literature (Hollenbach & Mckee, 1979; Lepp & Shull, 1983; Martin *et al.*, 1996; Galli & Palla, 1998; Tiné *et al.*, 1998; Le Bourlot *et al.*, 1999), nor that there is often substantial disagreement between the different rates.

Recently the situation has improved, thanks both to increasing computational power, and also to the development of more accurate interaction potentials, such as that of Boothroyd *et al.* (1996). The most accurate calculations at the present time are probably those of Le Bourlot *et al.* (1999). These involve a fully quantal treatment of non-reactive H–H_2 scattering (Flower, 1997; Flower & Roueff, 1998a), supplemented with an appropriate prescription for including the effects of reactive scattering. They also include the effects of collisions with He (Flower *et al.*, 1998), H_2 (Flower & Roueff, 1998b) and H^+ (Gerlich, 1990).

The calculations assume that the level populations are in statistical equilibrium, with negligible formation or destruction of H_2, and the resulting cooling rate is computed for a wide range of densities, temperatures, H_2 abundances and ortho-para ratios. The computed rates are available at **http://ccp7.dur.ac.uk/**.

2.5.2 Chemical changes

Changes in the chemical composition of the gas can lead to changes in its thermal energy. Energy is removed by endothermic reactions, such as collisional ionization, and produced by exothermic reactions, such as H_2 formation. In practice, however, the effects are usually small compared to the other sources of heating and cooling, and are included here primarily for completeness.

The rates all have the same basic form:

$$\Lambda_i = \Delta E_i k_i n_1 n_2. \tag{2.70}$$

The cooling rate Λ_i is given by the reaction rate multiplied by the net change in the thermal energy of the gas. In reactions such as collisional ionization, this is simply equal to the net change in binding energy; for example, H ionization leads to cooling at a rate

$$\Lambda = 2.176 \times 10^{-11} k_1 n_{\rm H} n_{\rm e} \,{\rm erg\,s^{-1}\,cm^{-3}}. \tag{2.71}$$

Other reactions that follow this pattern include the collisional dissociation of H^- (reactions 14,15 and 29) and of H_2 (reactions 12 and 13).

In reactions such as recombination, on the other hand, the energy released by the reaction is radiated away, and so does not alter the thermal energy. Moreover, the kinetic energy of the recombining electron is also radiated away. Recombination consequently leads to cooling, with ΔE in this case being given by the mean kinetic energy of the recombining electron. Cooling rates for the various reactions are given in Hui & Gnedin (1997). Reactions such as H^- formation (reaction 7) and H_2^+ formation (reaction 9) also lead to cooling in a similar fashion (Shapiro & Kang, 1987).

Finally, H_2 formation, whether via H^- or H_2^+, is exothermic, but deposits its energy into rotational and vibrational excitation of H_2, from where it is generally radiated away. It leads to significant heating only at high densities.

2.5.3 Bremsstrahlung

At temperatures greater than 10^6 K, primordial gas becomes fully ionized, and cooling by collisional excitation is no longer possible. At these temperatures, the cooling function becomes dominated by bremsstrahlung. At lower temperatures, it is of far less importance and plays no direct role in the cooling of most proto-galaxies. However, it does play an important role in the cooling of the hot gas produced by supernovae, and can be an important source of soft X-rays.

The frequency-averaged Gaunt factor appearing in the cooling rate is given by (Spitzer, 1978)

$$\langle g_{ff} \rangle = \begin{cases} 0.79464 + 0.1243 \log(T/Z^2) & (T/Z^2) < 3.2 \times 10^5 \text{ K} \\ 2.13164 - 0.1240 \log(T/Z^2) & (T/Z^2) > 3.2 \times 10^5 \text{ K} \end{cases} \quad (2.72)$$

2.5.4 Compton scattering

Free electrons within the gas will Compton scatter CMB photons, and will gain or lose energy depending upon whether the radiation temperature is higher of lower than the gas temperature. The net effect is to drive the gas temperature toward the radiation temperature. At high redshifts, this effect is very important, and keeps the IGM temperature closely coupled to the CMB temperature. However,

the rate falls off as z^4, being dependent on the radiation energy density, and below $z \sim 200$ it is no longer able to prevent the IGM from cooling adiabatically.

2.5.5 Photoionization and photodissociation

Photoionization causes heating because some fraction of the kinetic energy of the ejected photo-electron is inevitably transferred to the thermal energy of the gas. The heating rate is given by

$$\Gamma_{\rm ion} = 4\pi \int_0^\infty \frac{I_\nu \sigma_\nu}{h\nu} \eta(E) e^{-\tau_\nu} \, {\rm d}\nu, \tag{2.73}$$

for an isotropic radiation field, where σ_ν is the appropriate ionization cross-section, E is the kinetic energy of the photo-electron, and $\eta(E)$ is the energy-dependant heating efficiency. At low energies, $\eta(E) = 1$, but it falls off considerably once the electrons become energetic enough to cause secondary ionization. Values of η for a range of energies, gas compositions and fractional ionizations have been computed by Dalgarno *et al.* (1999); using their values, we obtain a heating rate that is accurate to within 25%.

Photodissociation of H_2 also heats the gas, with each dissociation typically producing 0.4 eV of heat (Black & Dalgarno, 1977). This gives a net heating rate of

$$\Gamma_{\rm pd} = 6.4 \times 10^{-13} k_{27} \, n_{\rm H_2} \ {\rm erg \ s^{-1} \ cm^{-3}}. \tag{2.74}$$

Additionally, photoionization of H_2, H_2^+ and H^- all heat the gas, but generally at rates that are too small to be important.

2.6 Summary

The approximations discussed in section 2.2 leave us with a simple chemical model. H_2 is formed via the molecular ions H^- or H_2^+ (reactions 7–10) and destroyed by collisions with H^+, e^- or H (reactions 11–13), by photoionization (reaction 24) or by photodissociation (reaction 27). H^- is formed by radiative association (reaction 7); as well as being destroyed during H_2 formation (reaction 8), it is also destroyed by collisional dissociation by electrons or neutral atoms (reactions 14,15 and 29), by reactions with H^+ ions (reactions 16 and 17) or by photodissociation (reaction 23). Similarly, H_2^+ is formed by radiative association

(reaction 9) and is destroyed by H_2 formation (reaction 10), by dissociative recombination (reaction 18) and by photodissociation (reactions 25 and 26). Since the formation of H^- and H_2^+ depend upon the electron and H^+ abundances respectively, we must also include the reactions governing the ionization of the gas – recombination (reactions 2,5 and 6), collisional ionization (reactions 1,3 and 4), photoionization (reactions 20–22) and charge transfer between hydrogen and helium (reactions 30 and 31).

In the absence of a radiation field, the accuracy of this model should be comparable to that of the Abel *et al.* (1997) model on which it is based, with the most significant reaction rates known to within 1% or better. The greatest source of uncertainty in this regime will come from the cooling function – the H_2 cooling rate has become increasingly well determined in recent years, but is unlikely to be as accurate as the chemical rates.

Including radiation does not have much effect on the accuracy of the model, as long as the gas is optically thin – the photodissociation cross-sections are known accurately, while the H_2 photodissociation rate is insensitive to details of the incident spectrum or H_2 level populations. In optically thick gas, on the other hand, the picture is not so rosy. Uncertainties in our treatment of secondary ionization limit the accuracy of the ionization rates to about 10 %, while the photodissociation rate may even more uncertain, unless the level populations are solved for explicitly. Nevertheless, this model should give a much better picture of the chemistry of optically thick primordial gas than previous treatments (Abel *et al.*, 1997; Kepner *et al.*, 1997). Moreover, as we shall see in chapter 4, the uncertainties in the chemical model ultimately have little effect on the results.

Chapter 3

Radiative feedback within the protogalaxy

3.1 Introduction

As outlined in chapter 1, primordial star formation is influenced by a variety of feedback mechanisms. Kinetic feedback, primarily in the form of supernovae, undoubtedly plays an important role in regulating primordial star formation, but a detailed understanding of its effects (as opposed to the simplified treatments of Mac Low & Ferrara 1999 or Ferrara & Tolstoy 2000, for example) is likely to necessitate the use of sophisticated numerical simulations and will be a long time in coming. It lies far beyond the scope of this thesis, and, although touched on briefly, it will not be discussed in detail.

Instead, I restrict my attention to radiative feedback. This can be usefully subdivided into *local* and *global* feedback. Global radiative feedback – the suppression of cooling within protogalaxies due to radiation from external sources – is discussed in the next chapter in the context of H_2 photodissociation. The other significant global feedback, reionization, occurs later than H_2 photodissociation (Haiman *et al.*, 2000) and in any case has already attracted a large amount of study (see Loeb & Barkana, 2001, and references therein).

This chapter is concerned with the local effects of radiative feedback – the way in which star formation within a protogalaxy affects gas cooling (and by implication star formation) within the *same* protogalaxy. In discussing this, I am primarily concerned with its effects on small, H_2-cooled protogalaxies, although the results presented will also apply to larger systems as long as the basic assumptions (*e.g.*

the lack of metals and dust) remain valid. I also restrict my attention to feedback occuring *before* the first generation of supernovae; the effects of radiative feedback at later times will be tied up with the effects of the supernovae themselves, and lie outside the scope of my study.

In practical terms, this means that the only sources of radiation we need consider are primordial stars; sources such as X-ray binaries or supernova remnants, which play an important role in global radiative feedback, post-date the first supernovae and can be ignored. Moreover, these primordial stars will be forming from unenriched primordial gas and will be metal-free. These stars – often known as population III – have properties that differ from those of their metal-enriched counterparts, as discussed in the next section.

Having discussed population III in section 3.2, I go on to investigate the effects of photoionizing radiation in section 3.3, while in section 3.4 I investigate the photodissociation of H_2 in a similar manner. I summarize my results in section 3.5.

3.2 Population III

The term population III has taken on a number of different meanings in the literature (Beers, 2000), but is used here to refer to *metal-free* stars, formed from unenriched primordial gas. Such stars, by definition, begin life with no carbon. [1] Their internal energy generation is dominated by the p-p chain rather than the CNO cycle, resulting in higher core temperatures for stars above 1 $M_\odot$; lower mass stars would in any case generate all of their energy via the p-p chain and are unaffected by the lack of carbon. These higher core temperatures in turn lead to higher effective surface temperatures, and hence harder spectra (Cojazzi *et al.*, 2000; Tumlinson & Shull, 2000).

Two groups have recently studied the production of ionizing photons by metal-free stars. Cojazzi *et al.* (2000) use the zero-metal stellar evolution models of Cassisi & Castellani (1993) and Forieri (1982) to construct appropriate isochrones. They then combine these with model atmospheres computed via the procedure of Auer & Heasley (1971) to obtain a set of spectra. Finally, they average over a Salpeter IMF (Salpeter, 1955), with minimum mass limit 0.024 $M_\odot$, to compute $N_{\rm ph}$, the

[1] Although massive stars may reach core temperatures high enough to form carbon via the triple-α reaction while still on the main sequence (Weiss *et al.*, 2000; Marigo *et al.*, 2001).

number of ionizing photons produced per second per solar mass of stars. They find that

$$N_{\rm ph} = 4 \times 10^{46}\ {\rm s}^{-1}\ {\rm M}_{\odot}^{-1}, \tag{3.1}$$

rising slightly as the population ages, until the most massive stars turn off the main sequence. This figure is around 25% higher than the rate at which ionizing photons are produced by an equivalent population of metal-poor stars.

Tumlinson & Shull (2000) follow a different approach. They use their own static stellar structure models to calculate luminosities, effective temperatures and surface gravities for stars over a wide range of masses. They then use the TLUSTY code (Hubeny & Lanz, 1995) to produce model atmospheres and associated spectra, from which $N_{\rm ph}$ can be calculated. They find that

$$N_{\rm ph} = 10^{47}\ {\rm s}^{-1}\ {\rm M}_{\odot}^{-1}, \tag{3.2}$$

differing by a factor of 2.5 from the results of Cojazzi *et al.*. This level of disagreement suggests that uncertainties remain in our knowledge of the basic stellar physics of these objects, and that both figures should therefore be treated with caution.

To reflect this uncertainty, let us write $N_{\rm ph}$ as

$$N_{\rm ph} = 10^{47} \chi\ {\rm s}^{-1}\ {\rm M}_{\odot}^{-1}, \tag{3.3}$$

where $\chi = 0.4$ for the Cojazzi *et al.* values, or $\chi = 1$ for the Tumlinson & Shull values.

The rate at which ionizing photons are produced by a stellar population of mass M_* solar masses can be written as

$$\begin{aligned} N_{\rm ion} &= M_* N_{\rm ph} \\ &= 10^{47} \chi M_*\ {\rm s}^{-1}. \end{aligned} \tag{3.4}$$

This can also be written in terms of the star formation efficiency. If we assume that all of the stars form in an instantaneous burst, then

$$N_{\rm ion} = 10^{47} \chi \epsilon f_b M\ {\rm s}^{-1} \tag{3.5}$$

where

$$\epsilon = \frac{M_*}{f_b M}; \tag{3.6}$$

here, $f_b \equiv \Omega_b/\Omega_m$ is the baryon fraction and M is the total mass of the protogalaxy. The star formation efficiency ϵ is determined by two separate quantities: the fraction of the baryons that are able to cool, and the efficiency with which this cool gas is converted into stars. Since neither of these numbers is well determined at present it is convenient to combine them into a single parameter.

These calculations assume that each ionizing photon is only responsible for a single ionization – in other words, that secondary ionization is negligible. This is a reasonable assumption, despite the harder spectrum of population III; any inaccuracy that it introduces will be small compared to the uncertainty in χ.

Turning to photodissociation, I first note that since it occurs via line absorption rather than continuum absorption, the number of photodissociating photons is not determined solely by the properties of the source. Any photon of sufficient energy can, in principle, cause the dissociation of an H_2 molecule, but the probability that one does so is determined both by its frequency and by the widths of the various Lyman-Werner lines. These in turn depend upon the state of the gas. Nevertheless, as I outline in section 3.4, these effects can be easily parameterised, at least in an approximate fashion, and it proves necessary to know the total number of photons that are energetically capable of photodissociating H_2. I take as appropriate energy limits $h\nu_1 = 11.15\,\mathrm{eV}$ and $h\nu_2 = 13.6\,\mathrm{eV}$. The lower limit corresponds to excitation from the para-hydrogen ground state to the least energetic level of the Lyman state. I ignore excitation from excited vibrational and rotational levels. The former assumption is justified by the populations of these levels; the latter can be justified by the fact that the dissociation probabilities corresponding to these transitions are very small. [2] The upper energy limit is simply the Lyman limit, higher energy photons being absorbed by the neutral hydrogen.

Neither Cojazzi *et al.* nor Tumlinson & Shull give values for the number of photons produced in the relevant energy range. However, inspection of their stellar spectra demonstrates that they are relatively featureless below the Lyman limit, and we can approximate them as black-bodies. This allows the number of photodissociating photons to be calculated easily.

If we denote the number of such photons produced per second by a star of mass

[2]Note that it is only transitions from excited rotational levels to the $v = 0$ level of the Lyman state that are omitted transitions to higher vibrational levels (which do have significant dissociation probabilities associated with them) fall within our energy range and are included.

M_s as N_s, then

$$N_s = \int_{\nu_1}^{\nu_2} \frac{L_\nu(M_s)}{h\nu}\, \mathrm{d}\nu. \tag{3.7}$$

Approximating the star as a black-body implies that

$$L_\nu = \frac{L_*}{\sigma T_{\rm eff}^4} \pi B_\nu, \tag{3.8}$$

where L_* is the total stellar luminosity, and B_ν is the Planck function

$$B_\nu = \frac{2h\nu^3}{c^2} \frac{1}{e^{h\nu/kT_{\rm eff}} - 1}. \tag{3.9}$$

Thus, N_s is given by

$$N_s = \frac{L_*}{\sigma T_{\rm eff}^4} \int_{\nu_1}^{\nu_2} \frac{\pi B_\nu}{h\nu}\, \mathrm{d}\nu. \tag{3.10}$$

Evaluation of this integral is straightforward.

Adopting values of L_* and $T_{\rm eff}$ as a function of stellar mass from Cojazzi *et al.* (2000), it is a simple matter to calculate N_s as a function of stellar mass. This is plotted in figure 3.1. If we average N_s over a Salpeter IMF, using the same minimum mass as Cojazzi *et al.*, we obtain $N_{\rm pd}$, the rate of photon production per solar mass

$$N_{\rm pd} = 3.7 \times 10^{45}\ \mathrm{s}^{-1}\ \mathrm{M}_\odot^{-1}. \tag{3.11}$$

The total number of photodissociating photons produced by a stellar population of mass M_* is then

$$N_{\rm dis} = 3.7 \times 10^{45} M_*\ \mathrm{s}^{-1}, \tag{3.12}$$

which we can link to the star formation efficiency of the protogalaxy as before

$$N_{\rm dis} = 3.7 \times 10^{45} \epsilon f_b M\ \mathrm{s}^{-1}. \tag{3.13}$$

Finally, note that these results do not include the effects of stellar evolution – as with ionizing photons, the change while stars remain on the main sequence is small, while details of post main sequence evolution remain poorly understood.

Another significant consequence of the metal-free nature of population III is the unimportance of stellar outflows. Stellar winds are ineffective without metal ions to drive them (Kudritzki, 2000; Hubeny *et al.*, 2000), while pulsation-driven outflows are less effective due to the increased stability of very massive metal-free stars compared to their metal-enriched counterparts (Baraffe *et al.*, 2001). Thus, outflows from population III stars will have a much smaller effect on their surroundings than those from stars in populations I or II.

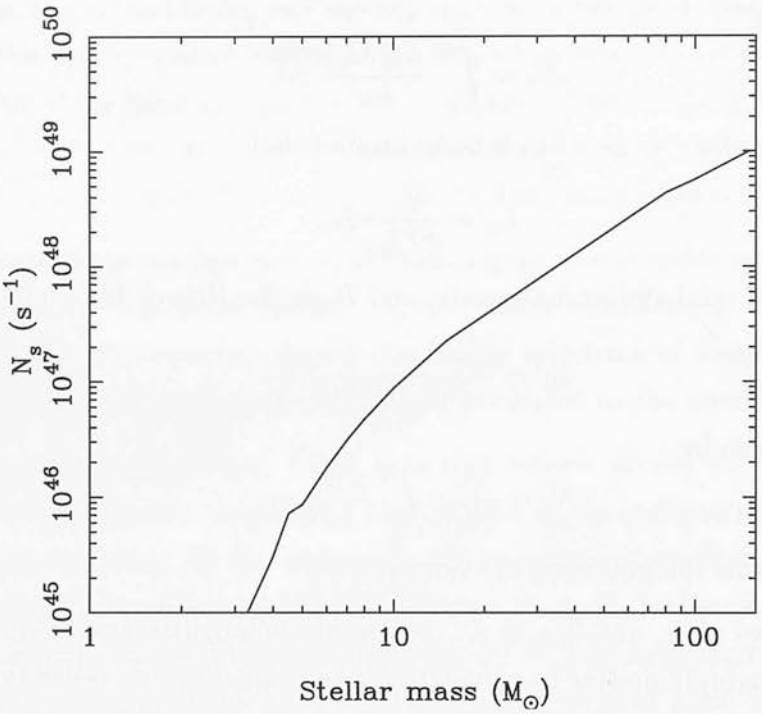

Figure 3.1: The number of photons with energies between 11.15 eV and 13.6 eV emitted per second by a metal-free star, as a function of stellar mass. The luminosities and effective temperatures of the stars are taken from Cojazzi *et al.* (2000)

3.3 Photoionization

Stars more massive than a few solar masses radiate a significant fraction of their energy above the Lyman limit and will quickly ionize the gas surrounding them, creating an H II region. Within this H II region, the gas temperature rapidly reaches equilibrium at $T \sim 10^4$ K and further cooling is prevented. It is reasonable to suppose that star formation within H II regions will be similarly suppressed, except within clumps dense enough to resist photoionization.

To determine the effectiveness of photoionization as a negative feedback process, we need to know how much gas will be incorporated into H II regions over the lifetime of the massive stars responsible for them. This in turn requires that we understand how H II regions evolve within a protogalaxy.

Below, I briefly describe the evolution of an H II region in a uniform density gas cloud of unlimited extent. This idealized model has been widely studied (see Yorke 1986 and references therein), and serves as a simple introduction to the

basic concepts. In section 3.3.2, I generalize the discussion to the case of a more realistic density profile – the truncated isothermal sphere model of Shapiro *et al.* (1999). Finally in section 3.3.3 I examine the fate of dense clumps of gas embedded in an H II region.

3.3.1 H II regions in uniform density gas

The simplest scenario involves a single source of ionization – a massive star or small stellar cluster – which switches on instantaneously inside a uniform density cloud of pure hydrogen. This immediately begins to ionize the surrounding gas, creating an H II region. The boundary of this H II region will have a thickness that is approximately equal to the mean free path of a photon at the ionization threshold. This is typically much smaller than any other length scale of interest and it is usually a good approximation to treat the H II region as being bounded by a sharp ionization front. In this case, the velocity of the front can be easily obtained: we equate the number of ionizing photons produced per second, $N_{\rm ion}$, to the number destroyed, both by the ionization of new material and by the need to balance recombinations within the ionized region.

Denoting the position of the ionization front by $r_{\rm I}$, we find that

$$\frac{{\rm d}r_{\rm I}}{{\rm d}t} = \frac{N_{\rm ion}}{4\pi n r_{\rm I}^2} - \frac{1}{n r_{\rm I}^2} \int_0^{r_I} r^2 n(r)^2 x^2 k_{2b} \, {\rm d}r, \tag{3.14}$$

where the fractional ionization $x \simeq 1$, and k_{2b} is the case B recombination coefficient for hydrogen (see table 2.1). Setting $x = 1$, and using the fact that n is constant, we can write this as

$$\frac{{\rm d}r_{\rm I}}{{\rm d}t} = \frac{N_{\rm ion}}{4\pi n r_{\rm I}^2} \left(1 - \frac{r_{\rm I}^3}{r_{\rm S}^3} \right) \tag{3.15}$$

where $r_{\rm S}$ is the Strömgren radius

$$r_{\rm S} = \left(\frac{3 N_{\rm ion}}{4\pi n^2 k_{2b}} \right)^{1/3} \tag{3.16}$$

at which the number of recombinations balance the number of ionizing photons.

Based on this, we can identify three distinct phases in the evolution of the H II region:

Phase 1: Initial growth

Initially, $r_{\rm I} \ll r_{\rm S}$ and recombinations can be neglected. In this case, [3]

$$\frac{{\rm d}r_{\rm I}}{{\rm d}t} = \frac{1}{3^{2/3}} \left(\frac{N_{\rm ion}}{4\pi n} \right)^{1/3} t^{-2/3}. \tag{3.17}$$

For a stellar cluster of mass $100\chi^{-1}\,{\rm M}_\odot$, we have $N_{\rm ion} = 10^{49}\,{\rm s}^{-1}$, and

$$\frac{{\rm d}r_{\rm I}}{{\rm d}t} = 4.5 \times 10^3 n^{-1/3} \left(\frac{t}{10^3\,{\rm yr}} \right)^{-2/3} \,{\rm km\,s^{-1}} \tag{3.18}$$

Thus, unless n is very large, the initial expansion of the ionization front will be highly supersonic. During this period it is known as an R-type front.

Phase 2: Pressure-driven expansion

So far, we have ignored any motion of the ionized gas. This is a good approximation as long as the front remains highly supersonic. However, after a time

$$\begin{aligned} t_{\rm ion} &= \frac{4\pi n r_{\rm S}^3}{3N_{\rm ion}} \\ &= (n k_{2b})^{-1} \\ &\simeq 1.2 \times 10^5 n^{-1}\,{\rm yr}, \end{aligned} \tag{3.19}$$

the front nears its Strömgren radius and its speed drops significantly. At this point, we can no longer ignore the thermal expansion of the ionized gas, which begins to drive a shock wave into the neutral gas ahead of the ionization front. Once this occurs, the ionization front is classified as D-type. Its speed is now subsonic with respect to the post-shock gas. The details of the transition from R to D-type front are rather complex and can probably best be followed by numerical simulation (Spitzer, 1978).

The subsequent expansion of the ionization front is due to the pressure-driven expansion of the H II gas. This decreases the density (and hence the recombination rate) within the H II region, allowing more H I to be ionized. During this expansion phase, a thin, dense H I shell forms between the shock and the ionization front. The radius of the front at a time t after the formation of the shock is given

[3] Note that our derivation does not take account of the finite speed of light and gives unphysically large values for the front velocity when t is of the order of the light-crossing time.

approximately by (Yorke, 1986)

$$r_{\rm I} = r_{\rm S}\left[1 + \frac{7}{4}\frac{c_{\rm s}t}{r_{\rm S}}\right]^{4/7} \tag{3.20}$$

where $c_{\rm s}$ is the isothermal sound speed within the H II region. The expansion phase continues until either the H II region attains pressure equilibrium with the surrounding gas or the ionizing source switches off.

Phase 3: Pressure equilibrium

Pressure equilibrium is attained at a radius

$$r_{\rm eq} \simeq \left(\frac{2T_2}{T_1}\right)^{2/3} r_{\rm S}, \tag{3.21}$$

where T_1 and T_2 are the temperatures of the H I and H II gas respectively. In a typical protogalaxy, $T_1 \sim 1000\,{\rm K}$ (or less), while the H II region has an equilibrium temperature $T_2 \sim 10^4$ K, implying that $r_{\rm eq} \sim 7r_{\rm S}$. Frequently, the lifetime of the massive stars powering the front will be too short to allow this equilibrium to be reached.

Inclusion of the effects of helium into this simple model changes the numerical results slightly, but otherwise has little effect.

Clumping

So far, our analysis has assumed that all of the gas is of uniform density. Inclusion of non-uniform clumping of the gas is relatively straightforward, as long as the mean density remains uniform and the clumping scale is much smaller than the size of the H II region. If these conditions are met, then we can simply replace n by the mean density $\bar{n}$, while n^2 in equation 3.16 must be replaced by $\overline{n^2}$. It is convenient to write this in terms of a clumping factor C, defined as

$$C = \frac{\overline{n^2}}{\bar{n}^2}, \tag{3.22}$$

so that r_S becomes

$$r_{\rm S} = \left(\frac{3N_{\rm ion}}{4\pi C\bar{n}^2 k_{2b}}\right)^{1/3}. \tag{3.23}$$

The initial growth and expansion phases proceed as before, but with this new value of $r_{\rm S}$.

3.3.2 H II regions in truncated isothermal spheres

A uniform sphere is a poor representation of a real protogalactic density profile, and thus the results of the previous section are of limited use. Rather, we require the analogous results for a more appropriate density profile.

At the outset, we assume that the density profile is spherically symmetric; relaxation of this assumption has little effect on the results unless the departure from spherical symmetry is large.

The simplest case that we can study is that of a power-law density profile. This is a reasonable representation of the results of numerical simulations (Fuller & Couchman, 2000; Abel *et al.*, 2000) if $n \propto r^{-2}$. The behaviour of H II regions in power-law profiles has been studied in detail by Franco *et al.* (1990) and I do not repeat their analysis here.

However, as discussed in section 1.3, uncertainty remains over the shape of the central portion of the density profile. In view of the sensitivity of the recombination rate to the density, it seems prudent to examine alternatives to the simple power-law model. One such alternative is the truncated isothermal sphere model of Shapiro *et al.* (1999). This has a density profile that is well approximated by

$$\begin{aligned} n(r) &= n_0\, n(\zeta) \\ &= n_0 \left(\frac{21.38}{9.08 + \zeta^2} - \frac{19.81}{14.62 + \zeta^2} \right) \end{aligned} \tag{3.24}$$

where $n_0 = 1.796 \times 10^4 n_{\mathrm{IGM}}$, and n_{IGM} is the density of the unperturbed IGM. The dimensionless radial coordinate $\zeta = r/r_0$, where r_0 is the core radius, given by

$$r_0 = \frac{63.67 h^{-2/3}}{1+z} \left(\frac{M}{10^6\ \mathrm{M}_\odot} \right)^{1/3} \mathrm{pc}. \tag{3.25}$$

For simplicity, I assume that the massive stars powering the H II region are located at the centre of the density profile, and that the cloud again consists of pure hydrogen. I look at the effects of relaxing the former assumption in due course; the latter merely changes the numerical results by a small amount, without significantly affecting the conclusions. As in the uniform density case, we can again divide the evolution of the H II region into three phases. However, as we shall see, the details of these phases are rather more complicated in the truncated isothermal sphere model.

ζ	$F(\zeta)$	$G(\zeta)$	$G(\zeta)/F(\zeta)$
0.25	5.14×10^{-3}	5.17×10^{-3}	1.01
0.5	3.96×10^{-2}	4.06×10^{-2}	1.03
1	0.276	0.303	1.10
2	1.37	1.89	1.38
5	3.99	11.1	2.79
10	4.79	24.7	5.15
15	4.94	35.2	7.13
20	5.00	44.5	8.91
25	5.03	53.3	10.6
29.4	5.05	60.7	12.0

Table 3.1: Selected values of the various dimensionless integrals that arise during the analysis of the growth of an H II region in a truncated isothermal sphere. $F(\zeta)$ is defined in equation 3.27; $G(\zeta)$ in equation 3.36.

Phase 1: Initial growth

Initially, the H II region will expand highly supersonically as an R-type front; its velocity is given by

$$\frac{\mathrm{d}r_\mathrm{I}}{\mathrm{d}t} = \frac{N_\mathrm{ion}}{4\pi n r_\mathrm{I}^2} - \frac{k_{2b} n_0 r_0^3}{r_\mathrm{I}^2} \frac{F(\zeta)}{n(\zeta)} \tag{3.26}$$

where $F(\zeta)$ is given by

$$F(\zeta) = \int_0^\zeta n(\zeta)^2 \zeta^2 \, \mathrm{d}\zeta, \tag{3.27}$$

and where we have again set $x = 1$. While $F(\zeta)$ cannot easily be reduced to simple analytical functions, it is trivial to evaluate numerically. In table 3.1 I list values of $F(\zeta)$ at selected points; note that since $\zeta_t = 29.4$ corresponds to the truncation radius of the sphere, $F(\zeta) \leq 5.05$.

The rapid expansion phase comes to an end when the ionization front nears its Strömgren radius. However, the finite size of the sphere means that this may never occur. If we assume that *all* of the gas in the sphere is ionized, then the total number of recombinations per second, N_rec, is given by

$$\begin{aligned} N_\mathrm{rec} &= \int_0^{r_t} 4\pi r^2 n(r)^2 k_{2b} \, \mathrm{d}r \\ &= 4\pi k_{2b} \, n_0^2 r_0^3 F(\zeta_t) \ \mathrm{s}^{-1}. \end{aligned} \tag{3.28}$$

If N_ion exceeds N_rec then there will never be enough recombinations to balance the production of ionizing photons. In this case, no Strömgren radius exists, and the front will continue to expand rapidly until the massive stars die.

Evaluating $N_{\rm rec}$, assuming a temperature of 10^4 K within the H II region, we find that

$$N_{\rm rec} = 3.5 \times 10^{48} (\Omega_b h)^2 \left(\frac{M}{10^6 \, {\rm M}_\odot} \right) (1+z)^3 \, {\rm s}^{-1} \tag{3.29}$$

which for a protogalaxy forming at $z = 25$ in our assumed cosmology becomes

$$N_{\rm rec} = 4.3 \times 10^{49} \left(\frac{M}{10^6 \, {\rm M}_\odot} \right) \, {\rm s}^{-1} \tag{3.30}$$

This calculation assumes that the gas is smoothly distributed. As before, it is straightforward to generalize to the case of clumpy gas provided that the clumping scale is small compared to the scale length of the protogalaxy. If this condition is met, then equation 3.29 becomes

$$N_{\rm rec} = 3.5 \times 10^{48} C (\Omega_b h)^2 \left(\frac{M}{10^6 \, {\rm M}_\odot} \right) (1+z)^3 \, {\rm s}^{-1} \tag{3.31}$$

where C is the clumping factor.

Equating this with equation 3.5 allows us to determine the star formation efficiency required for the H II region to be unbounded. We find that

$$\epsilon_{\rm crit} = 3.5 \times 10^{-5} \frac{C (\Omega_b h)^2}{f_b \chi} (1+z)^3. \tag{3.32}$$

For a protogalaxy at $z = 25$, and our assumed cosmology, this becomes

$$\epsilon_{\rm crit} = 3.5 \times 10^{-3} C \chi^{-1}. \tag{3.33}$$

Phase 2: Pressure-driven expansion

If $\epsilon < \epsilon_{\rm crit}$ then a Strömgren radius exists and the H II region is bounded. We can write the Strömgren radius as $\zeta_{\rm S} \equiv r_{\rm S}/r_0$. A closed form solution for $\zeta_{\rm S}$ cannot easily be given, but its value can be determined from

$$F(\zeta_{\rm S}) = \frac{N_{\rm ion}}{4\pi k_{2b} \, n_0^2 r_0^3}. \tag{3.34}$$

The ionization front reaches $\zeta_{\rm S}$ on a timescale

$$t_{\rm ion} = \frac{1}{n_0 k_{2b}} \frac{G(\zeta_{\rm S})}{F(\zeta_{\rm S})}, \tag{3.35}$$

where

$$G(\zeta_{\rm S}) = \int_0^{\zeta_{\rm S}} \zeta^2 n(\zeta) \, {\rm d}\zeta. \tag{3.36}$$

Values of $G(\zeta)/F(\zeta)$ are listed in table 3.1. As we might expect, $t_{\rm ion} \simeq (n_0 k_{2b})^{-1}$ in the (almost) constant density core, but it also remains a reasonably good approximation for larger H II regions. If we again assume a temperature in the ionized gas of 10^4 K, then we find that

$$t_{\rm ion} \simeq 4 \times 10^7 (1+z)^{-3} \, {\rm yr}, \tag{3.37}$$

corresponding to 10^3 – 10^4 yr for redshifts in the range 20 – 30. This implies that bounded H II regions will reach their Strömgren radius long before the end of the life of the massive stars powering them, requiring us to consider the next phase of their evolution.

As in the uniform density case, as the front nears $r_{\rm S}$ it changes from an R-type front to a D-type one as a shock front separates from it and precedes it into the surrounding gas. Subsequently, the expansion of the H II region is driven by the pressure of the hot, ionized gas.

At this point, we follow Franco *et al.* (1990) and make two significant approximations. We assume that the separation of the shock front and ionization front can be ignored; both are to be found at $r_{\rm I}$ and are separated by a shell of neutral gas of negligible thickness. This corresponds to what is known to happen in the uniform density case, and should be a reasonable approximation in the non-uniform case. Secondly, we assume that any density gradients within the H II region have been smoothed out, so that the ionized gas has uniform density. This is justifiable as long as the sound-crossing time in the H II region is short; it appears to be borne out by the results of numerical simulations (Franco *et al.*, 1990).

Making these approximations, we can write the mass of gas (both ionized and neutral) enclosed by the shock as

$$M_{\rm sh} = \int_0^{r_I} 4\pi m_{\rm H} n(r) r^2 \, {\rm d}r. \tag{3.38}$$

The mass of the ionized gas is given by

$$M_i = \frac{4\pi}{3} m_{\rm H} n_i r_I^3, \tag{3.39}$$

where n_i is the H^+ number density, and we have ignored the small contribution of the electrons towards the total mass. We can fix the value of n_i by requiring that ionizations balance recombinations within the H II region:

$$N_{\rm ion} = \frac{4\pi}{3} k_{2b} n_i^2 r_I^3, \tag{3.40}$$

and hence

$$M_i = \left(\frac{4\pi N_{\rm ion}}{3k_{2b}}\right)^{1/2} m_{\rm H} r_I^{3/2}. \tag{3.41}$$

Combining equations 3.38 and 3.41, we find that the ratio evolves as

$$\frac{M_{\rm sh}}{M_i} \propto \frac{G(\zeta_I)}{\zeta_I^{3/2}}. \tag{3.42}$$

Thus, the amount of mass in the neutral shell depends upon the quantity $G\zeta^{-3/2}$. If this increases as ζ increases, then the amount of mass in the neutral shell will also increase. On the other hand, if it *decreases* then the amount of neutral gas will likewise decrease. Should the amount of gas in the shell drop to zero, then this implies that that our model has broken down; the ionization front will overtake the shock, and will expand subsequently as an R-type front. Applying this to our truncated isothermal sphere model, we find that the ionization front overtakes the shock at a radius $\zeta_{\rm crit} = 4.7$.

There are thus three ways in which this phase of evolution could end: the massive stars could die, the H II region could reach pressure equilibrium, or the front could reach $\zeta_{\rm crit}$ and change back to an R-type front.

The first of these is unlikely to occur at high redshift; if we assume that the ionization front has a velocity equal to the speed of sound in the ionized gas, then it will reach the critical radius on a timescale

$$t_{\rm crit} = \frac{r_0}{c_s}\zeta_{\rm crit}. \tag{3.43}$$

For our canonical H II region temperature of 10^4 K, the sound speed is $11.4\,{\rm km\,s^{-1}}$ and

$$t_{\rm crit} = \frac{2.6 \times 10^7 h^{-2/3}}{1+z}\left(\frac{M}{10^6\ {\rm M}_\odot}\right)^{1/3}\ {\rm yr} \tag{3.44}$$

(where we have assumed the size of the initial Strömgren sphere to be negligible; clearly, the required time is smaller if it is not). For $z \sim 20 - 30$, we find that $t_{\rm crit} \simeq 1$ Myr and thus an H II region which reaches the pressure-driven expansion phase will generally also reach the next evolutionary phase before the massive stars powering it expire.

Phase 3: Pressure equilibrium or renewed expansion?

For the H II region to be in pressure equilibrium, it must satisfy the condition

$$n_i = \frac{T_1}{2T_2} n_1, \tag{3.45}$$

where n_1 is the pre-shock gas density, and T_1, T_2 are the temperatures in front of the shock and in the H II region respectively. Using equation 3.40 to write n_i in terms of the radius of the ionization front, we find that pressure equilibrium can be achieved if

$$n(r) > \frac{2T_2}{T_1}\left(\frac{3N_{\rm ion}}{4\pi k_{2b}}\right)^{1/2} r^{-3/2} \tag{3.46}$$

for some $r < r_{\rm crit}$. This condition can be re-written in terms of the initial Strömgren radius as

$$\zeta^{3/2} n(\zeta) > \frac{2T_2\sqrt{3F(\zeta_{\rm S})}}{T_1}. \tag{3.47}$$

Now, $\zeta^{3/2} n(\zeta) \leq 1.79$, implying that

$$F(\zeta_{\rm S}) < 0.27\left(\frac{T_1}{T_2}\right)^2. \tag{3.48}$$

If we have $T_1 = 10^3$ K and $T_2 = 10^4$ K as before, then a solution exists only if $F(\zeta_{\rm S}) < 2.7 \times 10^{-3}$. Inspection of table 3.1 shows that this requires the initial Strömgren radius to be well within the core radius of the isothermal sphere. Converting our constraint on $F(\zeta_{\rm S})$ into a constraint on $N_{\rm ion}$, we find that pressure equilibrium is possible only if

$$N_{\rm ion} \leq 1.9 \times 10^{45} (\Omega_b h)^2 \left(\frac{M}{10^6\ {\rm M}_\odot}\right)(1+z)^3\ {\rm s}^{-1}. \tag{3.49}$$

We can again use equation 3.5 to write this as a constraint on the star formation efficiency. We find that pressure equilibrium is possible only if

$$\epsilon \leq 1.9 \times 10^{-8} \frac{(\Omega_b h)^2}{\chi f_b}(1+z)^3 \tag{3.50}$$

which for our fiducial example of a protogalaxy at $z = 25$ in the usual cosmology becomes

$$\epsilon \leq 1.9 \times 10^{-6} \chi^{-1}. \tag{3.51}$$

This corresponds to a few solar masses of stars forming in a 10^6 M$_\odot$ protogalaxy, suggesting that the condition for pressure equilibrium will only rarely be met.

The effect of small-scale clumpiness of the gas depends upon the fate of the clumps. If they retain their identity during the initial growth and subsequent expansion phases, then equation 3.34 becomes

$$F(\zeta_{\rm S}) = \frac{N_{\rm ion}}{4\pi C k_{2b}\, n_0^2 r_0^3}, \tag{3.52}$$

and the constraint of $N_{\rm ion}$ is loosened by a factor of C. On the other hand, if they are rapidly smoothed out at the beginning of the expansion phase, then C quickly becomes one, and the above analysis applies.

Note that this calculation ignores pressure from non-thermal sources, such as ram pressure or turbulence that may serve to increase the effective external pressure. Therefore, we probably underestimate the required value of $N_{\rm ion}$. Nevertheless, it still does not seem likely that pressure equilibrium will occur in practice.

To recap, we have seen that there are three possible fates for an H II region within a truncated isothermal sphere:

1. If:
$$\frac{\epsilon}{(\Omega_b h)^2(1+z)^3} > 3.5 \times 10^{-5} C (f_b \chi)^{-1} \tag{3.53}$$
then no Strömgren sphere exists and the ionization front continues to expand as an R-type front for the whole of its life.

2. If:
$$3.5 \times 10^{-5} C (f_b \chi)^{-1} > \frac{\epsilon}{(\Omega_b h)^2(1+z)^3} > 1.9 \times 10^{-8} (f_b \chi)^{-1} \tag{3.54}$$
then the H II region initially expands rapidly, undergoes a transition to slow, pressure-driven expansion, but subsequently reverts to rapid expansion as the ionization front passes $\zeta_{\rm crit}$. The lower limit on $N_{\rm ion}$ in this inequality assumes that dense clumps are smoothed out within the H II region; if they are not, then it must be increased by a factor of C.

3. If:
$$\frac{\epsilon}{(\Omega_b h)^2(1+z)^3} < 1.9 \times 10^{-8} (f_b \chi)^{-1} \tag{3.55}$$
then the evolution of the H II region is much as it would be in the uniform density case: rapid expansion, followed by slow expansion, followed by pressure equilibrium. In practice, however, the constraint on the star formation efficiency is so small that it is almost certain to be violated.

Finally, note that we can put an upper limit on the amount of gas that has been ionized by the simple expedient of ignoring recombination. In this case
$$M_i(t) = 2.6 \times 10^{-50} N_{\rm ion} \left(\frac{t}{1 \text{ yr}} \right) \text{M}_\odot. \tag{3.56}$$

Off-centre sources

The rather lengthy discussion of the previous section assumes that the massive stars responsible for the H II region are located precisely at the centre of the density profile. In reality, this is unlikely to be the case. We therefore must consider how the results will change for a more realistic source distribution.

Moving the source away from the centre of the sphere introduces an angular dependence into the problem. There is no longer a single value for the Strömgren radius; rather, it depends upon the particular line of sight. Following Ricotti & Shull (2000), we can define the Strömgren radius along a given line of sight by

$$N_{\rm ion} = 4\pi k_{2b} \int_0^{r_S} r^2 n(R)^2 \, {\rm d}r, \tag{3.57}$$

where r is the radial distance from the source, and R is the radial distance from the centre of the sphere.

Clearly, if $N_{\rm ion}$ is large, then there may be no acceptable solution to this equation – the H II region may be unbounded. However, in contrast to the spherically symmetric case, there is no single critical value of $N_{\rm ion}$ at which the H II region becomes bounded. Rather, the value depends upon the line of sight, being largest for that which passes through the centre of the sphere, and smallest for that in the opposite direction. Thus, three types of behaviour are possible. If $N_{\rm ion} > N_{\rm max}$, defined by

$$N_{\rm max} = 4\pi k_{2b} \, n_0^2 r_0^3 H(\zeta_D), \tag{3.58}$$

where ζ_D is the distance of the source from the centre of the sphere, and

$$H(\zeta) = \int_0^{\zeta_D} \zeta^2 n(\zeta_D - \zeta)^2 \, {\rm d}\zeta + \int_0^{\zeta_t} (\zeta + \zeta_D)^2 n(\zeta)^2 \, {\rm d}\zeta, \tag{3.59}$$

then the H II region is unbounded in every direction, and will continue to grow for as long as the massive stars shine. On the other hand, if $N_{\rm ion} < N_{\rm min}$, where

$$N_{\rm min} = 4\pi k_{2b} \, n_0^2 r_0^3 \int_{\zeta_D}^{\zeta_t} (\zeta - \zeta_D)^2 n(\zeta)^2 \, {\rm d}\zeta \tag{3.60}$$

then the H II region is bounded in every direction. For intermediate values of $N_{\rm ion}$, the H II region will be bounded in some directions and unbounded in others, and will develop a very asymmetric shape (see, for example, figure 3 of Ricotti & Shull, 2000).

Assuming that the H II region is at least partially bounded, it will reach its Strömgren radius on a timescale $t_{\rm ion} = (n_0 k_{2b})^{-1}$. Its subsequent behaviour is complex, and ideally requires a numerical treatment. However, qualitatively it will be similar to the spherically symmetric case, with a period of pressure-driven expansion followed either by pressure equilibrium or by renewed supersonic expansion. The main difference to the spherical case is the dependence of the expansion rate on angle: the H II region will expand most rapidly down the steepest density gradient, *i.e.* radially outwards from the centre of the protogalaxy.

Even without a detailed numerical treatment, it is clear that one of the central results of this section – the fact that a substantial fraction of the protogalaxy will become ionized even if the star formation efficiency is low – must still hold. Moreover, the upper limit on the mass of ionized gas given by equation 3.56 also still applies.

3.3.3 The fate of dense clumps

Observations of local H II regions frequently show dense clumps of neutral gas embedded within them (see Elmegreen, 1998, and references therein). Many of these clumps are probably formed by dynamical instabilities at the edge of the H II region (Garcia-Segura & Franco, 1996), but it is possible that some pre-date the growth of the H II region and have survived the passage of the ionization front. Such clumps are natural sites for star formation, and the effectiveness of photoionization as a feedback mechanism will in large part depend upon the fate of these clumps.

A detailed study of the effects of ionizing radiation on dense, spherical clumps is presented in Bertoldi (1989) and Bertoldi & McKee (1990). They find that if the incident flux of ionizing photons is greater than

$$F_{\rm crit} = 2(k_{2b} r_c n_c^2 + c_s n_c)\ {\rm cm}^{-2}\ {\rm s}^{-1}, \tag{3.61}$$

where n_c and r_c are the clump density and radius respectively, then the clump will be ionized completely during the passage of the ionization front. The resulting ionized gas will have a significantly higher pressure than the surrounding inter-clump gas, and the clump will begin to expand, losing its identity within a few sound-crossing times. Clearly, star formation is suppressed in this case.

On the other hand, if $F < F_{\rm crit}$ then the passage of the ionization front does not

immediately ionize the clump. Rather, a D-type front (and associated shock) is driven into the clump, leading to significant compression of the gas. A fraction of the gas will be lost to photo-evaporation, but there is much evidence that the remaining gas will begin to form stars. [4] Thus, the growth of the H II region could actually lead to positive, rather than negative, feedback.

Adopting our usual H II region temperature of 10^4 K, we can write equation 3.61 as

$$F_{\rm crit} = 4.0 \times 10^{-2} M_c^{1/3} n_c^{5/3} + 2.3 \times 10^6 n_c \ {\rm cm}^{-2}\ {\rm s}^{-1}, \tag{3.62}$$

where M_c is the clump mass in units of $M_\odot$, and we have used the fact that $M_c = (4\pi/3) n_c r_c^3$ for a spherical clump of uniform density. For reasonable clump masses and densities, the second term dominates, and this reduces to

$$F_{\rm crit} \simeq 2.3 \times 10^6 n_c \ {\rm cm}^{-2}\ {\rm s}^{-1}. \tag{3.63}$$

Now, since

$$F_{\rm crit} = \frac{N_{\rm ion}}{4\pi R^2} \tag{3.64}$$

at a distance R from the source of the ionizing photons, we can use equation 3.63 to determine the critical distance at which clumps survive. Denoting this as $R_{\rm crit}$, we have

$$R_{\rm crit} = 190 \left(\frac{N_{\rm ion}}{10^{49}\ {\rm s}^{-1}} \right)^{1/2} n_c^{-1/2}\ {\rm pc}. \tag{3.65}$$

Thus, clumps of density $n_c \sim 10^3$ - 10^4 cm^{-3} or greater will generally survive within the H II region.

3.4 Photodissociation

As well as ionizing neutral hydrogen, massive stars will also photodissociate molecular hydrogen. As the H_2 is destroyed, the cooling time rises, until eventually the supply of cool gas ceases and star formation comes to a stop. It is clearly important to know how long this takes to happen, and how much gas it affects. Moreover, given that our understanding of protogalactic structure is still poor, it would be useful to be able to study this via simple analytical or semi-analytical methods, rather than by investing a large amount of time in detailed numerical modelling.

[4] See any of the references in table 1 of Elmegreen (1998).

If the gas was optically thin to photodissociating radiation, then investigating its effects would be easy. In the optically thin limit, the photodissociation rate depends simply upon the distance from the source, allowing a straightforward calculation of the evolution of the H_2 abundance and its final equilibrium value. The characteristic photodissociation timescale is easily obtained; if we assume that photodissociation is much quicker than H_2 formation or collisional dissociation, then it is simply

$$t_{\rm dis} = k_{27}^{-1} \simeq 7.2 \times 10^{-10} I(\bar{\nu})^{-1} \ {\rm s} \tag{3.66}$$

where $I(\bar{\nu})$ is the mean flux density in the Lyman-Werner bands, with units of erg s^{-1} cm^{-2} Hz^{-1} sr^{-1}. Omukai & Nishi (1999) have investigated this scenario and show that a single massive star can photodissociate H_2 throughout a small protogalaxy.

However, realistic protogalaxies are *not* optically thin. Including the effects of H_2 self-shielding is therefore necessary, but results in a far more complex problem – the photodissociation rate no longer depends simply on the distance to the source but also on the intervening H_2 column density. As H_2 is destroyed, this will change and thus the problem is inherently time-dependent; it becomes simple only once photodissociation equilibrium is reached.

Omukai & Nishi (1999) study the effects of self-shielding *assuming* that this equilibrium has already been reached. What is not clear from their analysis, however, is the time that it takes to reach this equilibrium. If this is substantially longer than lifetime of the massive stars, then the equilibrium solution derived by Omukai & Nishi, while perfectly correct, will never apply. It is therefore important to study the growth of the photodissociation region (PDR) prior to equilibrium.

We might hope to be able to do this in much the same way that we followed the growth of H II regions in the previous section. Unfortunately, this is not possible. Unlike an H II region, or indeed a PDR in a local molecular cloud, a PDR in a primordial protogalaxy will not be sharply bounded. This is a consequence of the low H_2 abundance, which causes the transition from optically thin to optically thick conditions to take place over a distance of the order of parsecs, compared to tenths or hundredths of a parsec for local PDRs. Thus, there is no well defined photodissociation front, and hence no easy way to apply the methods that we used for studying H II regions.

An alternative way of approaching the problem is to determine the timescale on which the amount of H_2 in the cloud as a whole significantly alters. As we shall see, an estimate of this can be obtained relatively easily via a simple analytical treatment.

Let us define the photodissociation timescale for a cloud containing a mass M_{H_2} of H_2 as

$$t_{\rm dis} = \frac{M_{H_2}}{\left|\dot{M}_{H_2}\right|}, \tag{3.67}$$

where $\dot{M}_{H_2}$ is the mass of H_2 destroyed per second.

We can write $\dot{M}_{H_2}$ as:

$$\dot{M}_{H_2} = -m_{H_2} N_{\rm dis} f_{\rm inc} f_{\rm abs} f_{\rm dis}. \tag{3.68}$$

where m_{H_2} is the mass of a H_2 molecule and $N_{\rm dis}$ is the rate at which photodissociating photons are produced (see section 3.2). A fraction $f_{\rm inc}$ of these photons are incident on the cloud, a fraction $f_{\rm abs}$ of the incident photons are absorbed and a fraction $f_{\rm dis}$ of these absorptions lead to dissociation.

The problem of calculating $t_{\rm dis}$ is reduced to that of calculating these few numbers. To do this precisely would involve a detailed treatment of radiative transfer through the cloud, but they are easy to estimate to a reasonable accuracy, allowing us to determine $t_{\rm dis}$ to within an order of magnitude. I outline the basis of these estimates in the following section.

3.4.1 Estimating the model parameters

$f_{\rm inc}$ The fraction of the emitted flux incident on a region of interest depends upon the size of the region and its position with respect to the star. For example, for a star in the centre of a spherically symmetric cloud of gas, clearly $f_{\rm inc} = 1$. On the other hand, for a star illuminating a compact cloud from a distance, $f_{\rm inc} = \Omega/4\pi$, where Ω is the solid angle subtended by the cloud, as seen from the star.

$f_{\rm abs}$ The fraction of incident photons absorbed in the cloud is obviously frequency dependent; far more photons are absorbed at frequencies corresponding to the centres of the Lyman-Werner lines than in the wings. However, rather

than tackle this frequency dependence directly, it is far simpler to approximate the effects of absorption in the set of Lyman-Werner lines by

$$f_{\rm abs} \simeq \frac{W}{W_{\rm max}}, \tag{3.69}$$

where W is the total dimensionless equivalent width of the set of lines and $W_{\rm max} = \ln\left(\frac{1110}{912}\right) \simeq 0.2$ is the dimensionless width of the range of wavelengths that contains all of the Lyman-Werner lines.

As in chapter 2, the equivalent width of a transition from a lower level l to an upper level u is given by

$$W_{ul}(N_l) = \int_0^\infty (1 - e^{-\sigma_\nu N_l}) \frac{{\rm d}\nu}{\nu}, \tag{3.70}$$

where N_l is the column density of absorbers in level l. To obtain the total equivalent width, we sum over the individual line widths as before, including a correction for line overlap:

$$W = \left(\frac{1 - e^{-w}}{w}\right) \sum_l \sum_u W_{ul}, \tag{3.71}$$

where $w = W_{\rm tot}/W_{\rm max}$, and

$$W_{\rm tot} = \sum_l \sum_u W_{ul} + \sum_{n=3}^{15} W_{\rm H}(1 \rightarrow n). \tag{3.72}$$

Data for H_2 is taken from Abgrall *et al.* (1993a,b), Roueff (1997) and Abgrall *et al.* (2000) as before; that for neutral hydrogen comes from Brocklehurst (1971). Voigt profiles are adopted for the absorption lines.

The resulting value of W depends upon the microphysics of the gas through Doppler broadening of the absorption lines. This can be parameterised in terms of the Doppler parameter

$$b = \sqrt{\frac{2kT}{m_{\rm H_2}} + v_{\rm t}^2}, \tag{3.73}$$

where $v_{\rm t}$ is the microturbulent velocity. Although the value of b is not known *a priori*, it can be constrained; it will lie between $b_{\rm min} = 1$ km s^{-1} (corresponding to pure thermal broadening at $T \sim 200$ K) and $b_{\rm max} = 10$ km s^{-1} (corresponding to thermal broadening at $T \sim 6000$ K, the temperature at which collisional dissociation of H_2 begins to dominate, plus a similarly sized contribution from

microturbulence). The effect of this uncertainty is greatest for column densities in the range $10^{15} < N_{H_2} < 10^{19}$ cm^{-2}, where Doppler broadening dominates the line profiles.

The relative populations of the ground state and the various excited rotational states also affect the value of W. Again, we do not know these populations *a priori* but we can constrain them; plausible values will lie somewhere between the LTE populations and the limiting case in which all of the molecules are in the para-hydrogen ground state.

Combining these two sources of uncertainty, we see that we can constrain f_{abs} by considering two limiting cases: one in which all of the H_2 lies in the para-hydrogen ground state, with Doppler parameter $b = b_{min}$, and another in which the levels have populations determined by LTE at a temperature of 6000 K, with $b = b_{max}$. In figure 3.2, I plot f_{abs} as a function of H_2 column density for these two cases. The limiting value of f_{abs} at high column density depends upon the H_2 abundance; the values plotted here assume an abundance $x_{H_2} = 10^{-4}$.

We see that our constraints are best in the low and high column density limits; in the Doppler-broadening dominated region $10^{15} < N_{H_2} < 10^{19}$ cm^{-2}, the uncertainty in f_{abs} can be as much as two orders of magnitude. However, it is worth remembering that these are fairly extreme cases and therefore provide strong constraints; realistic values of f_{abs} will generally lie near the middle of this range of values, and we would not expect the object-to-object variation to be particularly large.

f_{dis} The fraction of excitations that are followed by dissociation depends upon the vibrational state of the excited level; values for individual levels are given in Abgrall *et al.* (1992). Consequently, f_{dis} depends upon the shape of the incident spectrum and the column density of H_2. However, Draine & Bertoldi (1996) demonstrate that, for a power law input spectrum, the mean value of f_{dis} is 0.15; I adopt this value here. The error introduced by this approximation will be 20–30% at most, which is small compared with the uncertainty in f_{abs}.

I do not include the effects of re-emission of absorbed photons in the Lyman-Werner bands. As discussed in chapter 2, the effect on f_{dis} would be small – of the order of a few percent – which is less than the existing uncertainty.

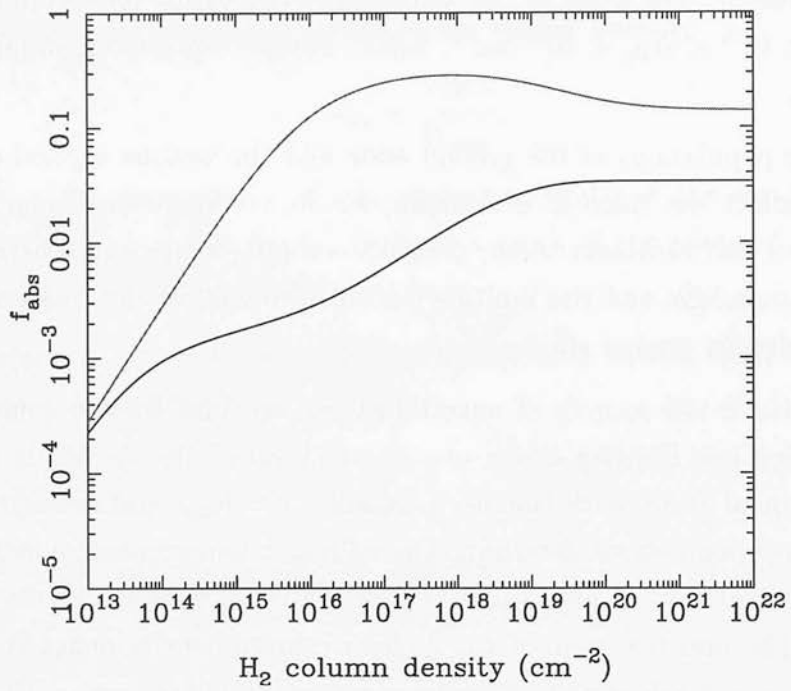

Figure 3.2: The fraction of radiation in the Lyman-Werner bands that is absorbed by the gas. The red line corresponds to H_2 with LTE level populations, and Doppler parameter $b = b_{\rm max}$; the black line to H_2 lying purely in the para ground state, with $b = b_{\rm min}$

3.4.2 Calculating the dissociation timescale

Putting all of these estimates together we have:

$$t_{\rm dis} = \frac{M_{\rm H_2}}{m_{\rm H_2} N_{\rm dis} f_{\rm inc} f_{\rm abs} f_{\rm dis}} \tag{3.74}$$

$$= 2 \times 10^{49} \frac{x_{\rm H_2} f_b M}{N_{\rm dis} f_{\rm inc} f_{\rm abs} f_{\rm dis}} \ {\rm yr}. \tag{3.75}$$

where $x_{\rm H_2}$ is the fractional abundance of H_2 [5] and M is the mass of the protogalaxy in units of $M_\odot$.

By using equation 3.13, this can also be written in terms of the star formation efficiency of the protogalaxy:

$$t_{\rm dis} = 5.4 \times 10^3 \frac{x_{\rm H_2}}{\epsilon f_{\rm inc} f_{\rm abs} f_{\rm dis}} \ {\rm yr}. \tag{3.76}$$

In order to arrive at our estimate for $t_{\rm dis}$, we have assumed that the formation of

[5] Which we assume to be constant throughout the protogalaxy.

H_2 can be neglected. This will generally be a reasonable approximation – near the star, H_2 formation is suppressed by the photodissociation of H^- and H_2^+, while further away the gas density is low and only a small fractional abundance can form within a massive star's lifetime.

3.4.3 Application to a protogalaxy

Having developed an approximation for $t_{\rm dis}$, the next step is to apply it to a model protogalaxy. As in section 3.3, I assume that the protogalaxy is spherically symmetric, with the source of Lyman-Werner photons located at or near the centre. Now, spherical symmetry implies that $f_{\rm inc} = 1$, and we know that $f_{\rm dis}$ is approximately constant, so we can reduce equation 3.76 to

$$t_{\rm dis} = 3.6 \times 10^4 \frac{x_{\rm H_2}}{\epsilon f_{\rm abs}} \ {\rm yr}. \tag{3.77}$$

We know that $x_{\rm H_2} > 10^{-4}$, or else the gas would not have cooled efficiently in the first place. On the other hand, as pointed out by Nishi & Susa (1999), recombination puts an upper limit on the H_2 abundance: $x_{\rm H_2} \lesssim 10^{-3}$, unless the gas is dense enough to form H_2 by three-body processes.

Of the remaining parameters, ϵ is truly independant, while $f_{\rm abs}$ depends upon the H_2 column density, and thus indirectly on the mass, formation redshift, H_2 abundance and density profile of the protogalaxy.

For a uniform sphere, the H_2 column density, measured radially from the centre to the edge, is

$$N_{\rm H_2} = 8.6 \times 10^{18} x_{\rm H_2} \left(\frac{M}{10^6 \ {\rm M}_\odot} \right)^{1/3} \left(\frac{\Omega_b}{\Omega_m^{1/3}} \right) h^{4/3} (1+z)^2 \ {\rm cm}^{-2}. \tag{3.78}$$

For a truncated isothermal sphere, on the other hand, it is an order of magnitude larger, being given by

$$N_{\rm H_2} = 9.7 \times 10^{19} x_{\rm H_2} \left(\frac{M}{10^6 \ {\rm M}_\odot} \right)^{1/3} \left(\frac{\Omega_b}{\Omega_m^{1/3}} \right) h^{4/3} (1+z)^2 \ {\rm cm}^{-2}. \tag{3.79}$$

As a fiducial example, let us consider a galaxy of mass $M = 10^6 {\rm M}_\odot$, H_2 abundance $x_{\rm H_2} = 5 \times 10^{-4}$ and formation redshift $z = 25$. In this case, in our assumed cosmology, $N_{\rm H_2} = 1.1 \times 10^{18} \ {\rm cm}^{-2}$ and thus

$$0.017 < f_{\rm abs} < 0.33. \tag{3.80}$$

The corresponding limits on the photodissociation timescale are

$$60\epsilon^{-1} < t_{\rm dis} < 1000\epsilon^{-1}\ {\rm yr}. \tag{3.81}$$

Thus, if $\epsilon \gtrsim 10^{-3}$, significant H_2 will be destroyed within the lifetime of the massive stars. This is an important result, as a star formation efficiency of 10^{-3} in a 10^6 $M_\odot$ protogalaxy corresponds to a mass of stars $M_* = 10^3 f_b \simeq 124$ $M_\odot$ and hence to only a few massive stars; it appears to confirm Omukai & Nishi's conclusion.

Nevertheless, it is important to examine how this conclusion depends upon the parameters we have assumed for the protogalaxy. The results of varying the parameters of our fiducial model are plotted in figures 3.3 – 3.5. In each case, $\epsilon = 10^{-3}$; as $t_{\rm dis} \propto \epsilon^{-1}$, it is trivial to rescale the results for a different star formation efficiency.

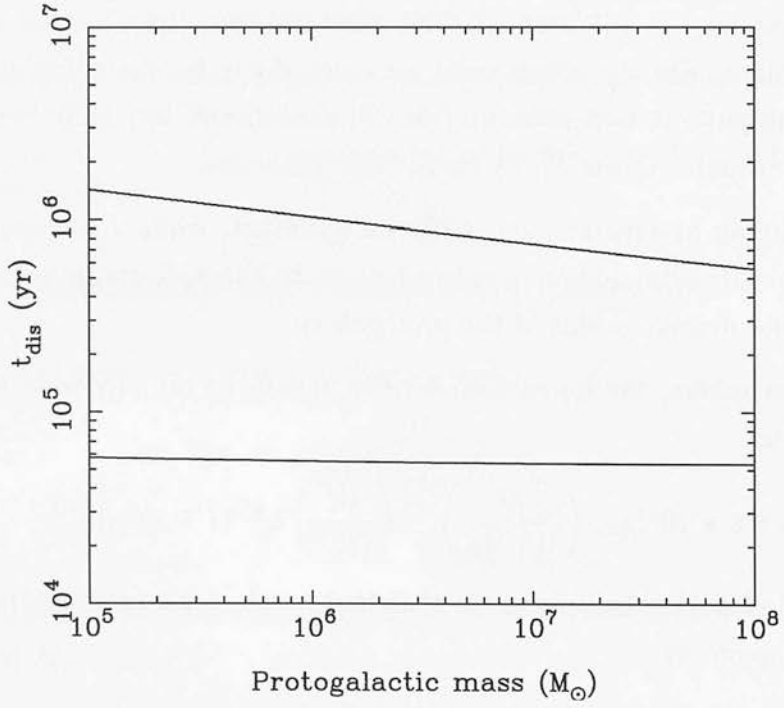

Figure 3.3: Upper and lower limits on the photodissociation timescale, plotted as a function of the mass of the protogalaxy. A formation redshift $z = 25$, H_2 abundance $x_{H_2} = 5 \times 10^{-4}$ and star formation efficiency $\epsilon = 10^{-3}$ are assumed.

In all three cases, the effects are small; we consistently find that

$$10^5 \lesssim t_{\rm dis} \lesssim 10^6\ {\rm yr} \tag{3.82}$$

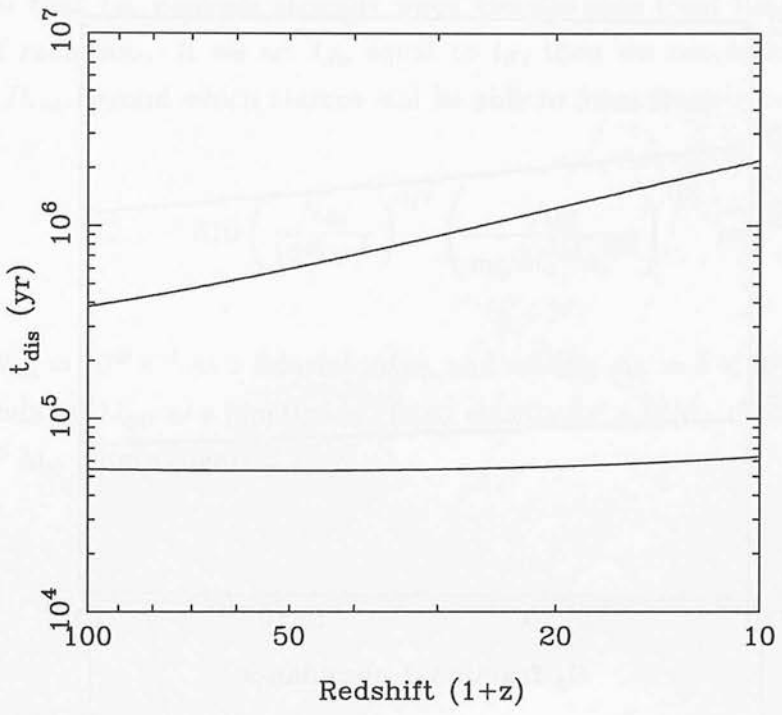

Figure 3.4: As figure 3.3, but examining the effects of varying the formation redshift. The protogalactic mass is $10^6 M_\odot$; other parameters are as before.

for $\epsilon = 10^{-3}$. Thus, even relatively inefficient star formation will produce enough radiation in the Lyman and Werner bands to destroy all of the H_2 within the protogalaxy within a short space of time.

3.4.4 Photodissociation and dense clumps

Our results above assume that the gas is evenly distributed. In reality, however, a significant fraction of the gas may be found in the form of dense clumps (Bromm *et al.*, 2002). If these clumps can withstand the effects of photodissociation for long enough, then they may be able to form stars. A proper treatment of the effects of photodissociative feedback must consider the fate of H_2 within such clumps.

If star formation is to occur within a clump, we would expect it to occur on a dynamical, or free-fall, timescale. The free-fall timescale, $t_{\rm ff}$, is given by

$$\begin{aligned} t_{\rm ff} &= \left(\frac{3\pi}{32G\rho}\right)^{1/2} \\ &\simeq 5 \times 10^7 n_c^{-1/2}\,{\rm yr}, \end{aligned} \tag{3.83}$$

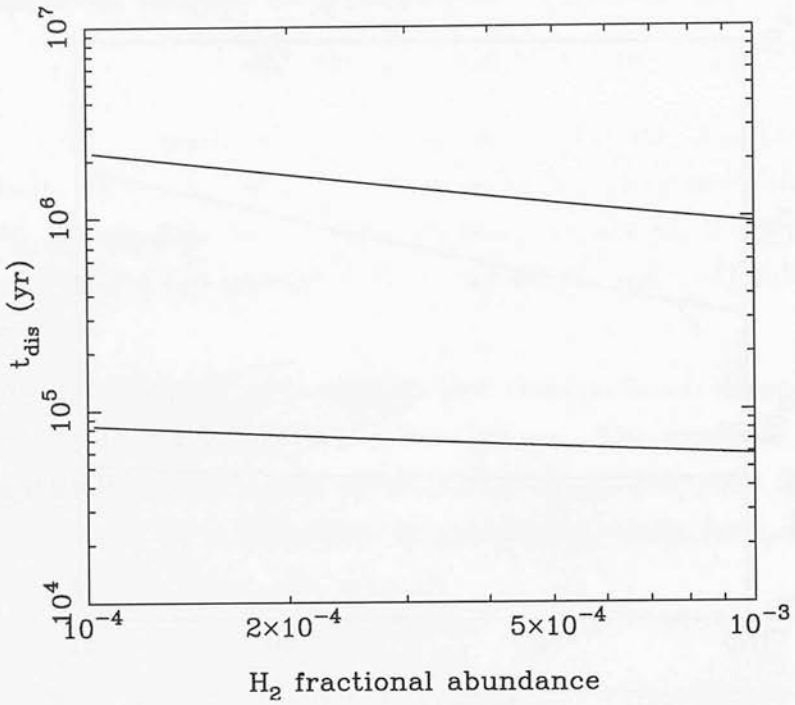

Figure 3.5: As figure 3.3, but examining the dependence upon the H_2 abundance. Again, a mass of $10^6 M_\odot$ is assumed.

(where we have assumed that the clump density, n_c, is uniform), and will typically be much smaller than the dynamical timescale of the protogalaxy. Star formation will only be suppressed if the photodissociation timescale is shorter than the free-fall timescale, *i.e.* $t_{\rm dis} < t_{\rm ff}$.

To calculate $t_{\rm dis}$ for a clump, we assume that it has uniform density; in this case, specification of the mass and density allows us to determine the clump radius, and hence $N_{\rm H_2}$ and $f_{\rm inc}$. The latter is given by

$$f_{\rm inc} = \frac{r_c^2}{4D^2}, \tag{3.84}$$

where r_c is the radius of the clump, D is the distance from the clump to the source of radiation, and we have assumed that the small-angle approximation is valid.

Substituting this into equation 3.75, and using our constant value for $f_{\rm dis}$, we find that

$$t_{\rm dis} = \frac{1.3 \times 10^{50}}{N_{\rm dis}} \frac{M_c^{1/3} n_c^{2/3} x_{\rm H_2}}{f_{\rm abs}} D^2 \ {\rm yr}, \tag{3.85}$$

where M_c is the clump mass (in units of $M_\odot$) and D is measured in parsecs.

It is clear that $t_{\rm dis}$ depends strongly upon the distance from the clump to the source of radiation. If we set $t_{\rm dis}$ equal to $t_{\rm ff}$, then we can identify a critical distance $D_{\rm crit}$ beyond which clumps will be able to form stars:

$$D_{\rm crit} = 610 \left(\frac{N_{\rm dis}}{10^{48}\ {\rm s}^{-1}} \right)^{1/2} \left(\frac{f_{\rm abs}}{x_{\rm H_2} M_c^{1/3} n_c^{7/6}} \right)^{1/2} \ {\rm pc}. \qquad (3.86)$$

Taking $N_{\rm dis} = 10^{48}\,{\rm s}^{-1}$ as a fiducial value, and setting $x_{\rm H_2} = 5 \times 10^{-4}$ as before, I have calculated $D_{\rm crit}$ as a function of clump density for a $10\,{\rm M}_\odot$ clump (figure 3.6) and a $10^3\ {\rm M}_\odot$ clump (figure 3.7).

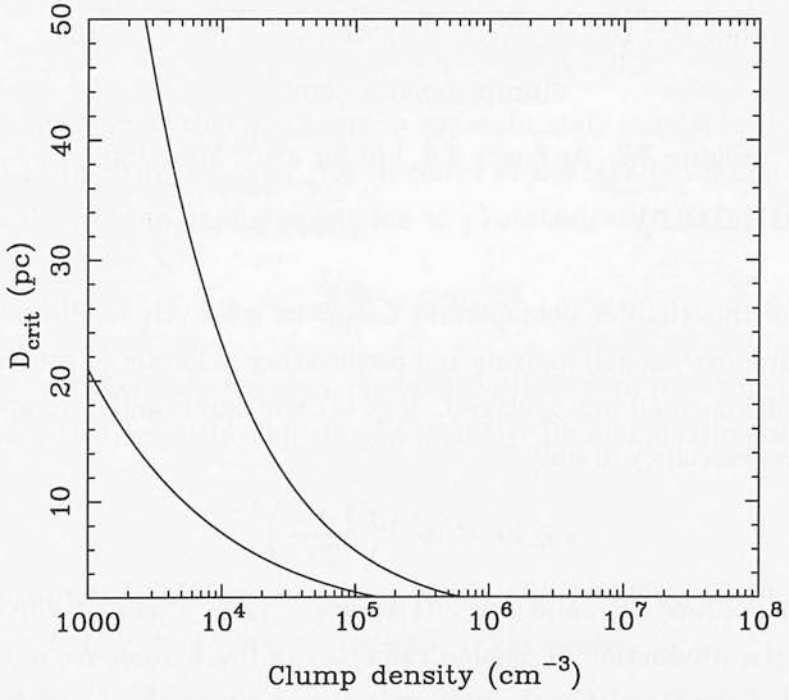

Figure 3.6: Upper and lower limits on $D_{\rm crit}$, as a function of clump density, for a clump of mass $10\ {\rm M}_\odot$.

The qualitative behaviour is clear. H_2 within high density clumps survives unless the clumps are very close to the massive stars, while H_2 in lower density clumps is destroyed out to considerable distances. The density at which the behaviour changes depends upon the clump mass and upon the photodissociating flux, but is typically of the order of $10^4 - 10^5\ {\rm cm}^{-3}$.

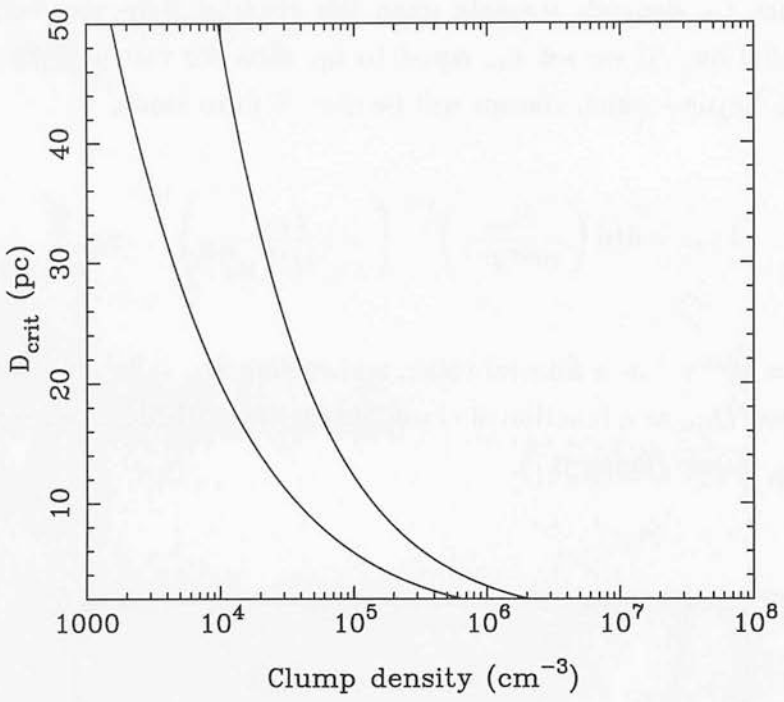

Figure 3.7: As figure 3.6, but for a 10^3 $M_\odot$ clump.

3.5 Summary

The results of this chapter demonstrate that even relatively inefficient star formation will produce enough ionizing and dissociating radiation to prevent further gas cooling within small protogalaxies. If gas is not significantly clumped then a star formation efficiency of only

$$\epsilon \gtrsim 1.6 \times 10^{-3} \left(\frac{1+z}{20} \right)^3 \tag{3.87}$$

is enough to produce an unbounded H II region, while a similar efficiency will also lead to the production of enough radiation in the Lyman-Werner bands to dissociate all of the H_2 within the protogalaxy on a timescale of less than a Myr.

At the same time, however, the results of sections 3.3.3 and 3.4.4 demonstrate that clumps of gas with densities greater than $n_{\rm crit} \sim 10^4$ cm^{-3} are not significantly affected by photoionization or photodissociation, and will continue to cool and to form stars.

Together, these results suggest that the protogalactic star formation efficiency is shaped in large part by hydrodynamical effects, being strongly dependent upon the fraction of gas that has been incorporated into dense clumps by the time

that the first massive stars form. This is not particularly well known, but high-resolution simulations such as those of Abel *et al.* (2000) or Bromm *et al.* (2002) are likely to make substantial progress in this area in coming years.

Finally, it is interesting to ask whether photoionization or photodissociation plays the greater role in suppressing gas cooling. One simple way in which we can compare their effects is to compare the mass of ionized gas with the mass of H_2 that has been destroyed. An upper limit on the former is given by

$$M_i(t) = 2.6 \times 10^{-50} N_{\rm ion} \left(\frac{t}{1 \text{ yr}} \right) \text{ M}_\odot \tag{3.88}$$

while the latter is given by the integral of equation 3.68, so that

$$\begin{aligned} M_{\rm dis}(t) &= m_{\rm H_2} N_{\rm dis} f_{\rm inc} f_{\rm abs} f_{\rm dis} t & (3.89) \\ &= 7.8 \times 10^{-51} N_{\rm dis} f_{\rm abs} \left(\frac{t}{1 \text{ yr}} \right) \text{ M}_\odot, & (3.90) \end{aligned}$$

where I have assumed that $f_{\rm abs}$ remains approximately constant. Using equations 3.5 and 3.13 to write $N_{\rm ion}$ and $N_{\rm dis}$ in terms of the star formation efficiency, we can write the ratio of photodissociation to photoionization as

$$\frac{M_{\rm dis}}{M_i} = 0.01 \frac{f_{\rm abs}}{\chi}. \tag{3.91}$$

It is clear that there are far more photoionizations than photodissociations, but then there is far more neutral hydrogen than H_2. Photoionization dominates only if

$$f_{\rm abs} < 10^2 \chi x_{\rm H_2}, \tag{3.92}$$

and thus is most important in protogalaxies with low H_2 column densities, becoming less significant as the column density increases.

Chapter 4

Global radiative feedback

4.1 Introduction

The previous chapter considered the effects of radiative feedback on small scales, and demonstrated that only a few massive stars are needed to suppress further star formation within a small protogalaxy. In this chapter, I turn my attention to the effects of radiative feedback on larger scales.

The basic principle is the same as on small scales – UV radiation can suppress star formation by destroying H_2 and by heating protogalactic gas. The difference is one of scale – when the opacity of the intergalactic medium (IGM) is low, gas in collapsing protogalaxies can be influenced by radiation from sources occupying a considerable fraction of the Hubble volume.

In this thesis, I restrict my discussion of radiative feedback to the period prior to reionization. Reionization itself has profound effects on star formation in low mass galaxies, but these effects are well known, and have attracted significant study (see, for example, Loeb & Barkana, 2001). Far less work has been done on the period preceding reionization. During this period, photoionization will clearly exert *some* negative feedback, but the high opacity of the IGM at the Lyman limit ensures that the effects of this are confined to regions near the sources; there is no global feedback.

Below the Lyman limit, however, the opacity of the IGM is very much lower and an alternative form of radiative feedback – photodissociation of H_2 by Lyman-Werner band photons – can operate. This form of feedback has the greatest effect on the smallest protogalaxies – those requiring H_2 in order to cool – and it is on

these that I concentrate. Its effects have been studied by a number of authors (Ciardi *et al.*, 2000a,b; Haiman *et al.*, 1996b, 1997, 2000; Machacek *et al.*, 2001).

The most self-consistent treatment is that of Haiman *et al.* (2000). Using a simple model for Lyman-Werner emission from protogalaxies, the details of which I discuss in section 4.2, they determine the spectrum and intensity of the resulting soft UV background as a function of redshift. They include the effects of intergalactic absorption by neutral hydrogen and H_2.

Since the global emissivity depends upon the star formation rate, and hence on the number of galaxies capable of forming stars, self-consistency requires that this calculation take into account the effect of the Lyman-Werner background on the galaxies. Accordingly, at each redshift, Haiman *et al.* calculate the critical temperature, $T_{\rm crit}$, at which a newly-formed protogalaxy exposed to the background is able to cool efficiently, using a method similar to that described in section 4.4. Having determined $T_{\rm crit}$, they can then calculate the background at a slightly lower redshift, and proceeding in this manner can track its evolution from high redshift until the onset of cosmological reionization.

Given reasonable assumptions concerning the star formation efficiency and primordial IMF, Haiman *et al.* find that a soft UV background inevitably develops prior to reionization, and that this background is strong enough to suppress H_2 cooling within newly formed protogalaxies. It causes $T_{\rm crit}$ to increase by over an order of magnitude to approximately 10^4 K, at which point Lyman-α cooling begins to dominate.

This is an important result, for a number of reasons. Firstly, it suggests that small protogalaxies do not play a significant role in reionizing the universe, even though the fraction of ionizing photons that can escape from them is far higher than in more massive galaxies (Ricotti & Shull, 2000). Moreover, if many of these small protogalaxies are prevented from forming stars then they may actually inhibit reionization. Without star formation to provide an internal source of ionization, they are effectively just large clouds of neutral gas and can act as significant sinks for ionizing photons from external sources, increasingly the clumping factor of the IGM and delaying recombination (Haiman *et al.*, 2001).

This result also suggests that small protogalaxies do not recycle a significant amount of gas into the IGM, and thus cannot explain the widespread metal enrichment seen in Lyman-α forest clouds (Cowie & Songaila, 1998; Ellison *et al.*,

2000). Finally, since most large galaxies that form in CDM models form from the merger of smaller galaxies, it is important to understand the history of these small galaxies properly in order to understand the initial conditions for the later stages of galaxy formation.

For all of these reasons, it is important to ensure that the Haiman *et al.* result is correct. However, they themselves identify at least one possible loophole in their analysis. Their simulations assume that the initial level of ionization in the gas corresponds to the residual level remaining after cosmological recombination, and that, following collapse, this subsequently decreases as the gas recombines. If there were some way to prevent recombination from occuring, however, and to maintain the fractional ionization at a non-equilibrium level then a substantially greater amount of H_2 could form. Increasing the H_2 abundance in this fashion would imply not only a greater cooling rate, but also a larger H_2 column density. This would increase the effectiveness of H_2 self-shielding and decreasing that of photodissociation. An enhanced level of ionization would at the very least reduce the effectiveness of negative feedback and may even overcome it altogether.

A plausible way in which an enhanced fractional ionization could be maintained is through the action of an early X-ray background. X-rays, unlike lower energy ionizing photons, are not absorbed strongly in neutral hydrogen or helium, and can propagate to large distances through the IGM. Thus, if any X-ray sources exist at these redshifts, they will naturally generate an X-ray background.

Haiman *et al.* (2000) briefly investigate the effects of such a background, within the context of a simple toy model. Their model assumes that the background takes the form of a power-law, $L_\nu \propto \nu^{-1}$, normalized such that its strength at the Lyman limit is a fraction $f_{\rm x}$ of the average background in the Lyman-Werner bands. The effects of absorption, whether in the IGM or intrinsic to the X-ray sources, are modelled as absorption by a fixed hydrogen column density $N_{\rm H} = 10^{22}\,{\rm cm}^{-2}$, plus a corresponding helium column density $N_{\rm He} = 0.08 N_{\rm H}$. Using this model, they find that one requires $f_{\rm x} \gtrsim 0.1$ in order to overcome negative feedback, and that for larger $f_{\rm x}$ cooling within the core region can actually increase.

The success of this model proves that, *in principle* X-rays can overcome negative feedback, but it is not (and was not intended to be) a realistic model of the X-ray background. In particular, it assumes a very high level of absorption. This may be appropriate for a background dominated by rare, luminous AGN, but it is far

from certain that AGN will play a significant role at high redshift – we know that their comoving number density decreases sharply at $z > 3$ (Pei, 1995), and there is already some evidence that they are not to be found in large numbers in the high redshift universe (Haiman *et al.*, 1999).

More plausible sources of X-rays are those associated with star formation, such as X-ray binaries or supernova remnants. These will be far more common than AGN, although individually less luminous, and the Haiman *et al.* toy model does not give an accurate picture of their effects. While one could clearly modify the model, the fact that these sources are directly associated with star formation allows their effects to be investigated in a far more self-consistent fashion. In the remainder of this chapter, I outline how this can be done, and the results that are obtained.

In section 4.2, I discuss how the emissivity (in the Lyman-Werner band and/or in X-rays) produced by a given source population can be calculated, and in section 4.3, I show how this emissivity can be used to determine the spectrum and intensity of the radiation background. In section 4.4, I outline the model used to investigate the effects of this background on protogalactic cooling, and in section 4.5 present a variety of results of this model. I conclude in section 4.6.

4.2 Modelling the emissivity

To determine the evolution of the UV or X-ray backgrounds, we first need to know how the emissivity, ε_ν, evolves. In general, the emissivity will depend upon both position and redshift. However, to simplify matters I assume that the emissivity (and hence the radiation field) is isotropic and homogeneous. This simplification is justified as long as the number of emitters is large (so that shot noise effects can be neglected) and if their typical clustering scale is much smaller than the scale over which they are visible. Generally, both of these conditions are satisfied for sources radiating in the Lyman-Werner bands or in the X-rays.

Since we are interested in sources of radiation which are directly associated with star formation, let us write the emissivity [1] as

$$\varepsilon_\nu = 1.04 \times 10^{-66} L_\nu \dot{M}_* \, \mathrm{erg\,s^{-1}\,cm^{-3}\,Hz^{-1}}, \tag{4.1}$$

[1] NB This is written in terms of proper coordinates; to convert to comoving coordinates, simply multiply by $(1+z)^3$.

where L_ν is the luminosity density per solar mass of stars formed per year, and $\dot{M}_*$ is the star formation rate per unit volume. [2]

Writing the emissivity in this manner means that we ignore any contribution from sources not directly related to the star formation rate, such as evolved stellar populations. This is a reasonable simplification, since Lyman-Werner emission will be dominated by emission from massive, short-lived O and B-type stars, while the X-ray luminosity of a star-forming galaxy is known to be correlated with its star formation rate (see section 4.2.3). The only significant sources of either type of radiation that do not fit this pattern are AGN, but we expect their influence to be negligible at very high redshift.

With this simplification, we can separate the problem of determining the emissivity into two conceptually independant portions: determining the star formation rate, and determining the luminosity density as a function of the star formation rate.

4.2.1 The global star formation rate

Although we have observational constraints on the star formation rate up to $z \sim 5$, we have no direct constraints (and few indirect ones) at higher redshift. Consequently, any model of the high redshift star formation rate will be entirely theoretical. Moreover, the lack of constraints suggests that we should choose as simple a model as possible. A good example is the star formation model adopted by Haiman *et al.* (2000), which is also used here. They assume that star formation proceeds primarily through starbursts, of duration $t_{\rm on}$ years, that are triggered when galaxies first form. During the starburst, the star formation rate is assumed to be constant. The cosmological density of star formation in this model is given by

$$\dot{M}_* = \frac{\epsilon f_b \Delta\rho_{\rm gal}}{t_{\rm on}} \, {\rm M}_\odot \, {\rm yr}^{-1} \, {\rm Mpc}^{-3}. \tag{4.2}$$

where ϵ is the star formation efficiency, $f_b \equiv \Omega_b/\Omega_m$, and $\Delta\rho_{\rm gal}$ is the cosmological density of matter in newly-formed galaxies (with units of ${\rm M}_\odot \, {\rm Mpc}^{-3}$). This can be written as

$$\Delta\rho_{\rm gal}(z) = \int_z^{z_1} \rho_m(z) \frac{{\rm d}F}{{\rm d}z}(z, T_{\rm crit}) \, {\rm d}z \tag{4.3}$$

[2] The former has units of erg ${\rm s}^{-1}$ ${\rm Hz}^{-1}$ $({\rm M}_\odot {\rm yr}^{-1})^{-1}$; the latter units of ${\rm M}_\odot$ ${\rm yr}^{-1}$ ${\rm Mpc}^{-3}$.

where $\rho_m = 2.8 \times 10^{11} \Omega_m (1+z)^3 \mathrm{M}_\odot \mathrm{Mpc}^{-3}$ is the cosmological matter density, and $F(z, T_{\rm crit})$ is the total fraction of matter in halos with virial temperatures greater than $T_{\rm crit}$. This fraction can be estimated from the Press-Schechter formalism:

$$F(z, M_{\rm crit}) = \int_{M_{\rm crit}}^{\infty} f(M, z)\, \mathrm{d}z \tag{4.4}$$

$$= \int_{M_{\rm crit}}^{\infty} \mathrm{erfc}\left[\frac{\delta_{\rm c}(z)}{\sqrt{2}\sigma(\mathrm{M})}\right] \mathrm{dz} \tag{4.5}$$

(see equation 1.126), where $M_{\rm crit}$ is the mass of a protogalaxy with virial temperature $T_{\rm crit}$. For a truncated isothermal sphere, this is

$$M_{\rm crit} = 1.5 \times 10^7 \left(\frac{T_{\rm crit}}{1000\ \mathrm{K}}\right)^{3/2} (1+z)^{-3/2} h^{-1}\ \mathrm{M}_\odot. \tag{4.6}$$

The upper limit of the integral in equation 4.3 is given by $z_1 = z + \Delta z(t_{\rm on})$, where $\Delta z(t_{\rm on})$ is the interval of redshift corresponding to the length of the starburst.

This model neglects any subsequent bursts of star formation, such as may be triggered by mergers, but as noted by Haiman *et al.*, the effect of gas recycling in mergers can be mimicked by choosing a large value of $t_{\rm on}$.

4.2.2 The UV luminosity density

The ultraviolet flux of a star-forming galaxy is dominated by emission from young, massive O and B-type stars. These are shortlived, with the most massive having lifetimes of only a few Myr, and thus the luminosity density will be strongly correlated with the star formation rate. Given enough information on the starbursts, such as the typical metallicity and IMF, it would be possible to use the techniques of stellar population synthesis (Bruzual & Charlot, 1993; Worthey, 1994) to determine the emitted flux. However, this would be an unnecessarily complicated approach for the simple investigation undertaken here. Instead, I use the well-known result that the emitted flux of a galaxy with constant star-formation rate is approximately flat between the Lyman and Balmer spectral breaks (Meier, 1976)

$$L_\nu = 2.6 \times 10^{27}\ \mathrm{erg\ s^{-1}\ Hz^{-1}}\ \left(\mathrm{M}_\odot\ \mathrm{yr}^{-1}\right)^{-1}. \tag{4.7}$$

Strictly speaking, this result applies only when the stellar population is in a steady state, with the rate at which massive stars form being balanced by the rate at which they die. However, given that we are summing over a large number of

different star-forming systems, each at a slightly different point in their evolution, it should be an adequate approximation.

4.2.3 The X-ray luminosity density

An empirical model

The correlation between the X-ray emission of star-forming galaxies and their star formation rate is well established. For example, Helfand & Moran (2001) collate data from a number of ASCA studies of local starbursts, and demonstrate that there is a linear correlation between the hard X-ray flux (measured in the 2 – 10 keV band) and the infrared flux [3] measured by IRAS (Sanders & Mirabel, 1996). Specifically,

$$F_{\rm X} \simeq 10^{-4} F_{\rm IR} \tag{4.8}$$

although a few sources, such as M82 (Moran & Lehnert, 1997) or NGC 3310 (Zezas *et al.*, 1998) have X-ray fluxes that are significantly larger. Similar correlations are found by David *et al.* (1992) for the 0.5 – 4.5 keV energy band, and Rephaeli *et al.* (1995) for the 2 – 30 keV energy band, although with constants of proportionality that differ by a factor of a few.

It is natural to assume that this correlation will continue to hold at high redshift. In this case, the X-ray luminosity can be written in terms of the star formation rate as

$$L_{\rm X} = 6 \times 10^{39} \left(\frac{\rm SFR}{1\,{\rm M_\odot\,yr^{-1}}} \right) \,{\rm erg\,s^{-1}}, \tag{4.9}$$

where we have used the result that

$$L_{\rm IR} \sim 1.5 \times 10^{10} \left(\frac{\rm SFR}{1\,{\rm M_\odot\,yr^{-1}}} \right) \,{\rm L_\odot}. \tag{4.10}$$

from the starburst model of Leitherer & Heckman (1995).

This value of $L_{\rm X}$ is an order of magnitude lower than that derived by Oh (2001) using a similar argument. Some of this discrepancy is due to the difference in the X-ray energy band considered (0.2 – 10 keV in Oh (2001), compared to 2 – 10 keV here); the rest is likely due to the intrinsic scatter in the observational data. With this in mind, we should clearly regard equation 4.9 as no more than an order of magnitude estimate of the X-ray luminosity.

[3] This is a good tracer of the star formation rate in dusty starbursts – see section 2.5 of Kennicutt (1998).

To convert the X-ray luminosity to the luminosity density, we need to know the shape of the X-ray spectrum. This frequently shows a great deal of structure (*e.g.* Dahlem *et al.*, 2000), but for our purposes can be adequately represented by a power law, $L_\nu \propto \nu^{-\Gamma}$. Rephaeli *et al.* (1995) find that a weighted average of their sample gives $\Gamma = 1.5 \pm 0.3$; I assume that this remains the case at high redshift. With this spectral index, the luminosity density becomes

$$L_\nu = 1.5 \times 10^{49} \nu^{-1.5} \text{ erg s}^{-1} \text{ Hz}^{-1} \left(\text{M}_\odot \text{ yr}^{-1}\right)^{-1} \tag{4.11}$$

which can be related to the emissivity as outlined above.

Alternative models

The above model has the benefits of being simple and empirically motivated. However, in the absence of observational confirmation, it remains an assumption and it is clearly prudent to consider alternatives. The simplest option is to modify the ratio of X-ray to infrared flux, or the spectral index, and in section 4.5 I examine the effects of making these changes.

More ambitiously, we might consider models in which L_ν is produced by a specific type of source, as this will give a better insight into how it may evolve with redshift.

Locally, X-ray emission from starburst galaxies is dominated by emission from massive X-ray binaries (David *et al.*, 1992). These consist of a massive O or B-type star in close orbit around a compact companion (a neutron star or black hole). Strong stellar winds from the massive star transfer mass onto the companion, powering a bright, but short-lived X-ray source. Emission from these sources is not directly dependent on redshift, but may be linked to metallicity, being enhanced in low-metallicity surroundings (van Paradijs & McClintock, 1995), although this is uncertain (Helfand & Moran, 2001). It requires the formation of a significant number of binary systems, as few will go on to form massive X-ray binaries. Locally, many stars are found in binary systems (Fischer & Marcy, 1992), but it is not clear that this will also be true at high redshift, particularly in metal-free gas, where numerical simulations find no evidence for fragmentation (Abel *et al.*, 2000).

As long as massive X-ray binaries dominate the X-ray emission, our empirical model remains a good one. Alternative models should therefore examine different

sources. Local observations suggest that the next most important X-ray sources will be supernova remnants (SNR). These can produce both thermal and non-thermal X-rays.

Their thermal X-ray emission is predominantly in the form of bremsstrahlung from hot gas in the remnant. Detailed modelling of this emission requires a hydrodynamical treatment of the evolution of the remnant (see *e.g.* Chevalier, 1999), but an order of magnitude estimate is easy to obtain. Helfand & Moran (2001) use the Sedov solution (Sedov, 1959) to estimate the X-ray emission, and find that for typical supernova parameters (an explosion energy $E \sim 10^{51}$ erg and an ambient density $n \sim 1\,\mathrm{cm}^{-3}$), a fraction $f_x = 2 \times 10^{-4}$ of the explosion energy is radiated at a typical temperature of 1 keV.

This estimate assumes that the supernova remnant reaches the Sedov-Taylor phase, but this need not be the case. If the ambient density is very high ($n \simeq 10^7\,\mathrm{cm}^{-3}$), then the supernova remnant will radiate its energy very rapidly, before the ejecta have had time to thermalize (Terlevich *et al.*, 1992). In this case, the X-ray luminosity is extremely large ($L_x \sim 10^{43}$ erg), but short-lived ($t \sim 1$ yr). The fraction of the explosion energy radiated as X-rays is significantly higher than in the standard case, with $f_x \simeq 0.01$, and the characteristic temperature is also much greater, being around 30 keV. At least one SNR of this type has been observed (Aretxaga *et al.*, 1999) and it is reasonable to expect them to be more common at high redshift, where the ambient densities will be greater and the effect of stellar winds less pronounced.

In section 4.5, I study two models in which the X-ray emission is produced by supernova remnants – one in which the emission comes only from the standard, low-luminosity remnants, and one in which it is all produced by high-luminosity, ultra-compact remnants. In both cases, the luminosity density can be written as

$$L_\nu = 1.3 \times 10^{26} \left(\frac{kT_x}{1\mathrm{keV}}\right)^{-1} \frac{E_{51}}{\nu_{\mathrm{sn}}} f_x e^{-h\nu/kT_x}\ \mathrm{erg\,s^{-1}\,Hz^{-1}}\ \left(\mathrm{M_\odot\,yr^{-1}}\right)^{-1}, \quad (4.12)$$

where E_{51} is the explosion energy in units of 10^{51} erg, T_x is the characteristic temperature of the emission and ν_{sn}^{-1} is the number of supernovae per solar mass of stars formed. The latter quantity depends on the IMF, and also the mass at which stars first become type II supernovae, but typically $\nu_{\mathrm{sn}}^{-1} \sim 0.01\,\mathrm{M_\odot^{-1}}$.

In addition to this thermal emission, supernova remnants also produce non-thermal X-rays. These are generated as the relativistic electrons produced in

the explosion lose energy, either through synchrotron radiation, or through inverse Compton scattering of infrared background photons. Synchrotron emission has been detected at low redshift (Koyama *et al.*, 1995; Petre *et al.*, 1999), but is generally not significant compared to the thermal emission. Inverse Compton (IC) emission has not yet been directly detected, but its presence has been invoked to explain the X-ray emission of M82 and NGC3256 (Moran & Lehnert, 1997; Moran *et al.*, 1999). At high redshift, we would expect inverse Compton emission to be far more significant – the energy loss rate of the relativistic electrons is proportional to the energy density of the CMB and thus scales as $(1+z)^4$. Oh (2001) has examined this scenario in some detail, and finds that IC emission exceeds synchrotron emission as long as the local magnetic field strength is less than

$$B < 1300 \left(\frac{1+z}{20} \right)^2 \mu\,\mathrm{G}. \tag{4.13}$$

This is large, even by the standards of the local universe, and it does not seem plausible that such strong magnetic fields could have been generated on small scales so early in the history of the universe.

The spectrum of inverse Compton emission depends upon the energy spectrum of the relativistic electrons, but is well approximated by $L_\nu \propto \nu^{-1}$ at the energies of interest. The intensity depends upon the fraction of supernova energy that is transferred to relativistic electrons. This is not well known, and may lie anywhere between 0.1% and 10%, corresponding to averaged X-ray luminosities in the range $3 \times 10^{38} - 3 \times 10^{40}\,\mathrm{erg\,s^{-1}}\,(\mathrm{M_\odot\,yr^{-1}})^{-1}$. I consider two models of inverse Compton emission, one corresponding to each end of this range of values. The resulting luminosity densities are given by

$$L_\nu = 1.86 \times 10^{38} \nu^{-1}\,\mathrm{erg\,s^{-1}\,Hz^{-1}}\,\left(\mathrm{M_\odot\,yr^{-1}}\right)^{-1} \tag{4.14}$$

and

$$L_\nu = 1.86 \times 10^{40} \nu^{-1}\,\mathrm{erg\,s^{-1}\,Hz^{-1}}\,\left(\mathrm{M_\odot\,yr^{-1}}\right)^{-1} \tag{4.15}$$

respectively. These estimates assume a high-energy cut-off of 10keV, but are only logarithmically dependent on the value of this cut-off.

4.3 Calculating the background

Once we have determined the redshift evolution of the emissivity, the next step is to examine the radiation background that it produces. If absorption is negligible, then determining this background is simple. The mean specific intensity of the background at an observed frequency ν_0 and redshift z_0 is given by (Madau *et al.*, 1999)

$$J(\nu_0, z_0) = \frac{1}{4\pi} \int_{z_0}^{\infty} \varepsilon_\nu(z) \frac{(1+z_0)^3}{(1+z)^3} \frac{c}{H(1+z)} \, \mathrm{d}z, \tag{4.16}$$

where $\nu = \nu_0(1+z)/(1+z_0)$. The background is completely determined by the emissivity – given a model for the evolution of $\varepsilon_\nu(z)$, calculation of $J(\nu_0, z_0)$ is trivial.

If absorption is *not* negligible, then equation 4.16 must be modified to include its effects; we can write

$$J(\nu_0, z_0) = \frac{1}{4\pi} \int_{z_0}^{\infty} \varepsilon_\nu(z) \frac{(1+z_0)^3}{(1+z)^3} \frac{c}{H(1+z)} e^{-\tau} \, \mathrm{d}z, \tag{4.17}$$

where $\tau(\nu_0, z_0, z)$ is the optical depth at frequency ν_0 due to material along the line of sight from redshift z_0 to z.

There are two distinct contributions to τ: intrinsic absorption (*i.e.* absorption by material within the galaxy that contains the source of radiation), and absorption by material in the intergalactic medium. Denoting these as τ_{int} and τ_{IGM} respectively, we can write the total optical depth as

$$\tau = \tau_{\mathrm{int}} + \tau_{\mathrm{IGM}}. \tag{4.18}$$

4.3.1 Intrinsic absorption

Intrinsic absorption is difficult to model with any degree of accuracy as it will depend upon a number of variables – the size and shape of the galaxy, its ionization state, the position of the sources within it, the dust content *etc.* Rather than attempt to model these in detail – a significant undertaking in itself – I instead assume that it can be approximated by absorption by a neutral hydrogen column density of $N_{\mathrm{H}} = 10^{21}\ \mathrm{cm}^{-2}$, plus a neutral helium column density of $N_{\mathrm{He}} = 8 \times 10^{19}\ \mathrm{cm}^{-2}$. This is assumed to be constant, and to be the same for all sources. With this assumption, τ_{int} becomes

$$\tau_{\mathrm{int}} = 10^{21}\sigma_{20}(\nu) + 8 \times 10^{19}\sigma_{21}(\nu) \tag{4.19}$$

for X-rays; for Lyman-Werner photons, I take it to be zero since any H_2 in the host galaxy will be rapidly destroyed [4] This model for intrinsic absorption is similar to that used by Haiman *et al.* for the total absorption in their toy model of the X-ray background, although the smaller value of $N_{\rm H}$ adopted here is more appropriate for X-ray sources that are not AGN.

4.3.2 Absorption in the IGM

Compared to intrinsic absorption, the effects of absorption in the IGM are relatively straightforward to determine. We can write $\tau_{\rm IGM}$ as

$$\tau_{\rm IGM}(\nu_0, z_0, z) = \int_{z_0}^{z} \kappa(\nu, z) \frac{c}{H(1+z)} \, {\rm d}z, \tag{4.20}$$

where $\kappa(\nu, z)$ is the absorption coefficient. The form of κ depends upon the type of absorption – whether continuum or line based – which differs for the two types of radiation.

X-ray absorption

For X-rays, the opacity is dominated by continuum absorption by neutral hydrogen and helium, as the He^+ abundance is small prior to reionization. This allows us to write $\tau_{\rm IGM}$ as

$$\tau_{\rm IGM}(\nu_0, z_0, z) = \int_{z_0}^{z} \left[\sigma_{20}(\nu) n_{\rm H}(z) + \sigma_{21}(\nu) n_{\rm He}(z)\right] \frac{c}{H(1+z)} \, {\rm d}z, \tag{4.21}$$

where $\nu = \nu_0(1+z)/(1+z_0)$ as before and where $n_{\rm H}(z)$ and $n_{\rm He}(z)$ are the neutral hydrogen and helium densities in the IGM. Assuming that most of the gas in the IGM remains smoothly distributed, and that the ratio of hydrogen to helium retains its primordial value then this becomes

$$\tau_{\rm IGM}(\nu_0, z_0, z) = \int_{z_0}^{z} \left[0.927\sigma_{20}(\nu) + 0.073\sigma_{21}(\nu)\right] \frac{n_b(z) c}{H(1+z)} \, {\rm d}z, \tag{4.22}$$

where n_b is the baryon number density. Determination of $\tau_{\rm IGM}$ by numerical integration of this equation is straightforward.

[4] See chapter 3.

Lyman-Werner band absorption

Lyman-Werner band photons face two main sources of opacity. The first is absorption by the Lyman series lines of neutral hydrogen. In a neutral IGM, these lines have very large optical depths, and absorb virtually all of the Lyman-Werner photons redshifted into them. As demonstrated by Haiman *et al.* (2000), this causes the spectrum of the Lyman-Werner background to develop a 'sawtooth' shape, an example of which is displayed in figure 4.1. This 'sawtoothing' is due to the fact that, at a frequency ν and redshift z_0, sources are visible only as far as a maximum redshift $z_{\rm max}$, given by

$$\frac{1+z_{\rm max}}{1+z_0} = \frac{\nu_{\rm l}}{\nu}, \tag{4.23}$$

where $\nu_{\rm l}$ is the frequency of the nearest, higher-energy Lyman series line. The nearer ν is to $\nu_{\rm l}$, the nearer $z_{\rm max}$ is to z_0, and the smaller the number of sources that are visible. Consequently, at observed frequencies just below a Lyman series line, very few sources are visible, while at frequencies just above the line, a large number can be seen. Hence, the observed spectrum declines steadily until the line is reached, and then jumps sharply up, giving it its characteristic sawtooth shape.

The other source of opacity in the IGM comes from absorption by the Lyman-Werner lines of intergalactic H_2. If we assume that none of the absorbed photons are re-emitted, then the optical depth due to H_2 becomes

$$\tau_{\rm IGM}(\nu) = 2.654 \times 10^{-2} \sum_i f_{\rm osc,i} \lambda_i \frac{n_{\rm H_2}(z_i)}{H(z_i)} \tag{4.24}$$

where we sum over all lines with frequencies ν_i that lie between ν and $\nu_{\rm l}$; z_i is the redshift at which line i is seen in absorption, and is given by $(1+z_i)/(1+z_0) = \nu_i/\nu$. For simplicity, I have approximated the lines as delta functions; comparison with the results of Haiman *et al.* (2000), who use Voigt profiles, shows this to be a good approximation. At typical IGM number densities, the population of excited states will be negligible, while work by Flower & Pineau des Forêts (2000) shows that the ortho-para ratio will be approximately 0.25. An example of the resulting opacity is plotted in figure 4.2.

In reality, of course, some of the absorbed photons will be re-emitted. Approximately 85% of absorptions do not result in dissociation, with roughly 5% resulting in re-emission of the original Lyman-Werner photon, while about 15% result in

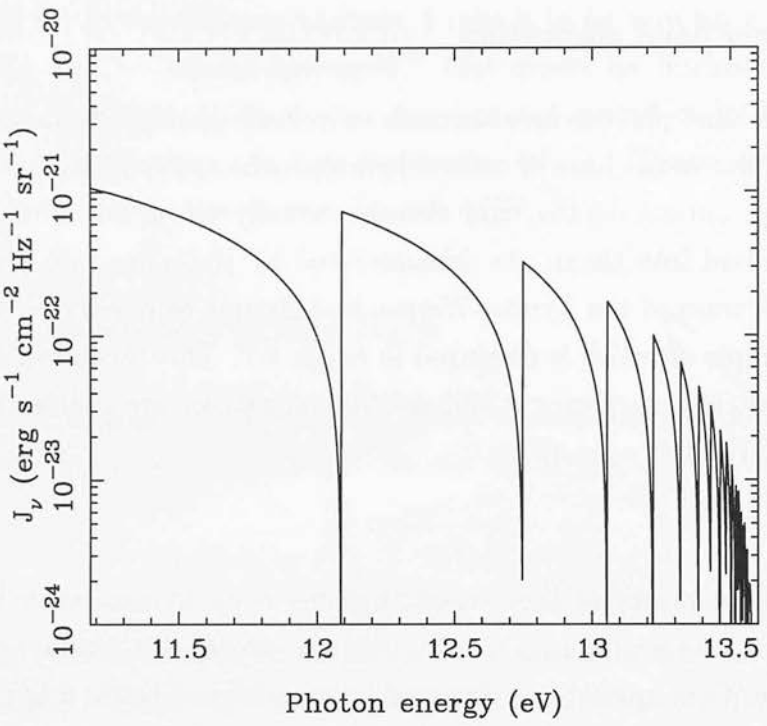

Figure 4.1: An example of the 'sawtooth' pattern that neutral hydrogen absorption creates in the spectrum of the Lyman-Werner background. This example shows the background at $z = 30$, and is taken from a model with star formation efficiency $\epsilon = 0.1$ and starburst lifetime $t_{\rm on} = 10^7$ yr.

the emission of a photon elsewhere in the Lyman-Werner band system (Glover & Brand, 2001). Moreover, although many of these photons will be scattered out of the line of sight, the assumed isotropy of the radiation field implies that this will be balanced by an equal number of scatterings into the line of sight. Equation 4.24 is thus correct only to within about 20%.

A more accurate treatment of H_2 opacity would clearly be desirable, but actually proves to be relatively unimportant. This is because the magnitude of the Lyman-Werner background required to rapidly destroy intergalactic H_2 is far less than that required to significantly affect protogalactic cooling. This is obvious from the results of Haiman *et al.* (2000), and can also be verified numerically. Let us suppose, for simplicity, that there is no X-ray background, and that H_2 formation can be neglected. In this case, the H_2 abundance is governed by

$$\frac{{\rm d}n_{\rm H_2}}{{\rm d}t} = -k_{27} n_{\rm H_2} \tag{4.25}$$

$$= -1.38 \times 10^9 \, J_\nu n_{\rm H_2}, \tag{4.26}$$

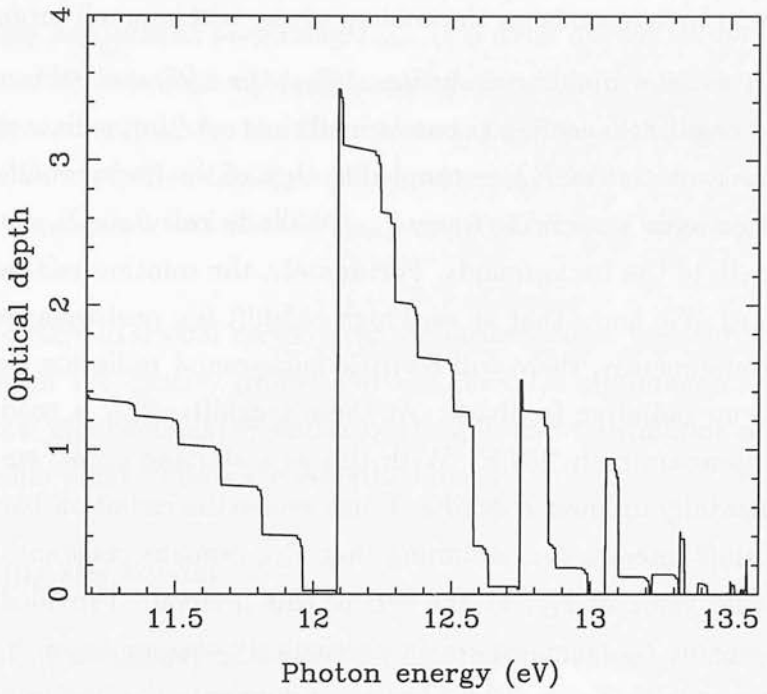

Figure 4.2: The optical depth from intergalactic H_2 at $z = 30$. I assume an H_2 abundance of 2.2×10^{-6} (Stancil *et al.*, 1998) and neglect the destruction of intergalactic H_2 by the Lyman-Werner background.

and H_2 dissociates on a timescale

$$t_{\rm dis} = 7.25 \times 10^{-10} J_\nu^{-1} \, {\rm s}. \tag{4.27}$$

This becomes shorter than the Hubble time if

$$\begin{aligned} J_\nu &> 3.5 \times 10^{-27} h \sqrt{\Omega_m} (1+z)^{3/2} \, {\rm erg \, s^{-1} \, cm^{-2} \, Hz^{-1} sr^{-1}} \\ &\gtrsim 1.1 \times 10^{-23} \left(\frac{1+z}{20} \right)^{3/2} \, {\rm erg \, s^{-1} \, cm^{-2} \, Hz^{-1} sr^{-1}} \end{aligned} \tag{4.28}$$

(where in the second line I have adopted my usual cosmological model). Once J_ν exceeds this value, intergalactic H_2 is rapidly destroyed. This level of flux, however, has only a marginal effect on cooling within protogalaxies, suggesting that absorption by intergalactic H_2 is of little importance to the evolution of $T_{\rm crit}$.

4.4 Simulating protogalactic cooling

The preceding sections have shown how, given simple models of star formation and X-ray emission, we can calculate the spectrum and intensity of the Lyman-

Werner and X-ray backgrounds. In this section, I describe how we can investigate the effect of these backgrounds on the cooling of gas within small protogalaxies.

Our main aim is to determine the evolution of $T_{\rm crit}$, the critical virial temperature below which protogalactic cooling becomes inefficient. An immediate problem is the fact that the evolution of $T_{\rm crit}$ is coupled to that of the backgrounds – to calculate the backgrounds we need to know $T_{\rm crit}$, while to calculate $T_{\rm crit}$ we need to know the strength of the backgrounds. Fortunately, the solution to this problem is straightforward. We know that at very high redshift few protogalaxies will yet have formed; consequently, there will be little background radiation (aside from the CMB) and no radiative feedback. At these redshifts, $T_{\rm crit}$ is readily determined, and is approximately 250 K. With this as a starting point, we can then proceed incrementally to lower redshifts. I first evolve the radiation backgrounds for a small redshift interval Δz, assuming that $T_{\rm crit}$ remains constant, and then calculate the true value of $T_{\rm crit}$ at the end of this interval. Provided that Δz is small, the error in $T_{\rm crit}$ remains small, particularly once emission from larger protogalaxies (ones with $T_{\rm vir} > 10^4$ K) begins to dominate the background. This strategy reduces the coupled problem to the simpler one of determining $T_{\rm crit}$ given a radiation background; in the remainder of this section I outline my method for doing this

Ideally, one would use a detailed hydrodynamical simulation to calculate $T_{\rm crit}$, as in Machacek *et al.* (2001). This is plausible if the background radiation is negligible or if the protogalactic gas is optically thin, as the photochemical reaction rates can be calculated in advance. Once the gas becomes optically thick, however, this method becomes impractical, as to obtain the photoionization and photodissociation rates we must solve for the radiative transfer of the ionizing or dissociating photons. This adds greatly to the computational cost of the problem, ruling out any systematic investigation. Since the protogalaxies in which we are interested will be optically thick, this approach is not appropriate.

An alternative approach, used by Haiman *et al.* (2000), is to simplify the problem by not attempting to simulate the hydrodynamical evolution of the protogalaxy. Rather, we assume that the density profile remains static, at least on the timescale of the problem, and solve for the radiative transfer of ionizing and photodissociating radiation within this static profile. Further simplification can be had by assuming that this profile is spherically symmetric.

This is clearly a dramatic approximation, but, as we saw in chapter 1, it proves surprisingly accurate at predicting $T_{\rm crit}$. It is of no use for studying protogalactic evolution after the onset of strong cooling, but this is outside the scope of our study. Moreover, it allows the thermal and chemical evolution of a model protogalaxy to be computed rapidly – in a matter of minutes on a fast workstation – allowing the redshift evolution of $T_{\rm crit}$ to be studied for a large number of different source models.

The basic computational method can be broken down into three portions – initialization of the density profile and the chemical abundances, computation of the thermal and chemical evolution of the gas, and termination of the simulation at a suitable point. These are described below.

Initializing the model

I model the protogalaxy as a truncated isothermal sphere (Shapiro *et al.*, 1999), with central density

$$n_0 = 1.796 \times 10^4 n_{\rm IGM}, \tag{4.29}$$

(where $n_{\rm IGM}$ is the density of the unperturbed IGM), and truncation radius

$$r_t = 1.874 \times 10^3 \left(\frac{M}{10^6\ {\rm M}_\odot} \right)^{1/3} (1+z)^{-1} h^{-2/3}\ {\rm pc}. \tag{4.30}$$

We can use the relationship between mass and virial temperature for a truncated isothermal sphere (equation 4.6) to write the latter as

$$r_t = 4.62 \times 10^3 \left(\frac{T_{\rm vir}}{1000\ {\rm K}} \right)^{1/2} (1+z)^{-3/2} h^{-1}\ {\rm pc}. \tag{4.31}$$

The density profile is thus completely specified by the virial temperature and redshift of formation.

This profile is subdivided into 100 spherical shells of uniform thickness – experimentation has shown that this is a sufficient number to properly resolve the protogalaxy. Once the gas density in each shell has been calculated, the initial number densities of the various chemicals are set. The initial chemical abundances are taken from Stancil *et al.* (1998), with the ortho to para-hydrogen ratio coming from Flower & Pineau des Forêts (2000). At low redshifts, these abundances will be inaccurate, as intergalactic H_2 is destroyed by the Lyman-Werner background, but the results of the simulations prove to be insensitive to the precise values of the initial abundances.

Computing the evolution

Since the gas density is fixed, we can write the equations for the evolution of the chemical abundances in the form (Anninos *et al.*, 1997)

$$\frac{\mathrm{d}n_i}{\mathrm{d}t} = C_i(n_j, T) - D_i(n_j, T)n_i \tag{4.32}$$

where n_i is the number density of a chemical species i, C_i is the sum of all of the processes that create species i and D_i is the sum of all of the processes that destroy it. Moreover, the temperature equation can be written in a similar fashion

$$\frac{\mathrm{d}T}{\mathrm{d}t} = \frac{2}{3k_b n}\left[\Gamma(n_j, T) - \Lambda(n_j, T)\right], \tag{4.33}$$

where Γ and Λ are the total heating and cooling terms respectively. Together, these equations form a coupled set of ordinary differential equations that must be solved simultaneously.

In optically thin gas (or when the background radiation is negligible), the thermal and chemical evolution of each shell is independent of that of the others. In this case, it is a simple matter to compute the evolution of each shell. The only difficulty is presented by the stiff nature of the rate equations. This rules out the use of explicit integration methods; instead, I use the implicit STIFBS routine of Press *et al.* (1986). Even with an implicit integration method, chemicals with small characteristic timescales present a problem. In the optically thin case, these problems manifest only for the two molecular ions, H^- and H_2^+, which both have equilibrium timescales of the order of 10 – 100 yr. It is possible, although inconvenient, to solve the rate equations using timesteps of this size, as long as we can calculate the photochemical rates in advance. Once we add in the overhead of recalculating them regularly, as we must in optically thick gas, such small timesteps become impractical. Accordingly, I approximate by assuming that the H^- and H_2^+ abundances both come instantaneously to equilibrium. This introduces some uncertainty into the calculated H_2 abundance, but this is of the same order of magnitude as the H^- and H_2^+ abundances, and hence is completely negligible.

In optically thick gas, the problem is far more complicated. The photochemical rates can no longer be computed in advance; rather, they will depend upon the radiation field seen by each shell, which itself depends upon the opacity of each of the shells. As the chemical abundances evolve, these opacities change, causing

the photochemical rates to change. Thus, the chemical evolution of any one shell will generally depend upon that of all of the other shells.

My treatment of this problem is similar to that of Kepner *et al.* (1997). I begin by noting that the photochemical rates at a distance r from the centre of the protogalaxy will depend upon the mean flux $J_\nu(r)$ at that point. This can be written in terms of the isotropic external radiation field, $J_\nu(r_t)$, as

$$J_\nu(r) = \frac{1}{2} J_\nu(r_t) \int_{-1}^{+1} \exp\left[-\tau(\nu, r, \mu)\right] \, \mathrm{d}\mu \tag{4.34}$$

where $\mu = \cos\theta$ and θ is the angle between the line of sight and the radial direction. This expression is valid as long as the re-emission of photons within the gas can be ignored; as we saw previously, this is a good approximation.

Above the Lyman limit, we can write the opacity as

$$\begin{aligned} \tau(\nu, r, \mu) &= \sigma_{20}(\nu) N_{\mathrm{H}}(r, \mu) + \sigma_{21} N_{\mathrm{He}}(r, \mu) & (4.35) \\ &= \left(\sigma_{20} + 0.08\sigma_{21}\right) N_{\mathrm{H}}(r, \mu) & (4.36) \end{aligned}$$

where I have assumed that the He^+ abundance is negligible. Photodissociation can be treated in a similar manner, since the photodissociation rate for any rotational level is a monotonic function of the column density of H_2 in that level. Our radiative transfer problem therefore reduces to one of calculating column densities.

Accurate calculation of the H_2 photodissociation rate does, however, present a significant problem. As we saw in chapter 2, the photodissociation rate in optically thick gas depends upon the populations of the rotational levels of H_2. Consequently, rather than simply solving for the H_2 abundance, we should calculate the abundance of H_2 in each of its rotational levels (or at least the subset of them which can acquire significant populations). In principle, this is straightforward, but in practice the short lifetimes of the excited states are a real problem, as they again force us to solve the chemical equations using very small timesteps. This makes it impractical to calculate the chemical evolution for any significant length of time. Clearly, some approximation is necessary.

An approximation like that used for H^- and H_2^+ initially seems promising, but is difficult to implement – collisional excitation and de-excitation couple the various levels together, requiring us to solve a sizeable set of simultaneous equations for

the individual level populations for each timestep. I have chosen to use instead a couple of simpler approximations

The first of these approximations ignores excited H_2 altogether; we assume that all of the H_2 is in the $J = 0$ or $J = 1$ rotational states, and that the ortho-para ratio has its equilibrium value of three to one. As we saw in chapter 2, this is a better approximation than simply assuming that all of the H_2 is para-hydrogen, and probably underestimates the actual photodissociation rate by no more than a factor of two to three at most.

The other approximation takes the opposite tack; we assume that the levels are in local thermodynamic equilibrium, in which case their populations depend solely on the temperature. Together, these two approximations bracket the true level populations. Comparison of the results of simulations run with the different approximations shows us how much effect our uncertainty has on the calculated values of $T_{\rm crit}$. Selected results of such simulations are presented in section 4.5; briefly, I find that although in some cases there is a significant difference in the evolution of the H_2 abundance, the evolution of $T_{\rm crit}$ is broadly similar, and we can use our simple approximation – that there is no excited H_2 – with reasonable confidence.

Returning to my discussion of the computational method, recall that the radiative transfer problem can be reduced to one of calculating column densities, from which we can then calculate the mean flux by use of equation 4.34. For each grid point, therefore, we need to calculate the column density along a suitable number of lines of sight. Obviously, the better the angular resolution, the smaller the error in $J_\nu(r)$; on the other hand, there is little point in being unnecessarily precise. Having run trial simulations with various different angular resolutions, I find that splitting the integral into 100 equally spaced bins proves adequate.

Thus, for each shell we must calculate the column densities of atomic and molecular hydrogen along 100 lines of sight. These column densities are given by

$$N_i(r,\mu) = \int_0^L n_i[r'(x)]\,\mathrm{d}x \tag{4.37}$$

where $L = r\mu + [r_t^2 - r^2(1-\mu^2)]^{1/2}$ is the distance to the edge of the protogalaxy and $r'(x) = (r^2 + x^2 - 2rx\mu)^{1/2}$ is the distance to the centre of the protogalaxy as a function of x, the position along the line of sight. Since the shells are evenly spaced, it is simple to reduce this integral to the appropriate sum.

Once we have the column densities and the mean flux, we can then calculate the various photochemical rates. In practice, since the $J_\nu(r_l)$ is assumed to be constant on these timescales, we can do most of this computation in advance, storing the results in a series of look-up tables, and interpolating rates as necessary. This inevitably introduces a small amount of error, but this can be made negligible by using sufficiently fine divisions in our tables.

Finally, one last approximation is necessary, since it is impractical to recalculate the column densities and photochemical rates for every single timestep taken by STIFBS. Rather, I recalculate them on the timescale on which *they* vary. The procedure is outlined below:

1. Choose an appropriate starting timestep, Δt. I usually err on the side of caution, and use an initial timestep of $\Delta t = 10^{10}$ s.
2. Calculate the column densities, and hence the photochemical rates.
3. Evolve the model protogalaxy for Δt, assuming that the photochemical rates remain constant over this period. Store the results.
4. Return to the beginning of the timestep, and evolve the model protogalaxy again, this time using *two* timesteps, each of half the size.
5. Compare the results of the two simulations. If any of the chemical abundances or temperatures have relative errors of more than 10^{-3}, then reject the results of the timestep, and start again from step one with a new timestep of half the size.
6. If we accept the results, then update the starting time from t to $t + \Delta t$, and choose a new timestep. This choice is guided by the size of the discrepancy between the two runs – if it is very small, then I enlarge the timestep, if not then I keep it at its current size. This is a fairly crude form of adaptive stepsize control, but works well in practice.

Stopping the simulation

Using the method outlined above, I run the simulation until one of two criteria is met: either we exceed a time limit, $t_{\rm lim}$, or the protogalaxy begins to cool strongly.

A time limit is necessary in order to take account of the fact that these protogalaxies are not completely isolated objects and will only survive for a limited period before merging. The distribution of survival times can be calculated from the Press-Schechter formalism (Lacey & Cole, 1993), but for reasonably rare objects the mean survival time is typically of the order of the Hubble time. Accordingly, I set $t_{\rm lim} = t_H$.

The second criterion – the onset of strong cooling – can be assessed in a number of ways, as we saw in chapter 1. Haiman *et al.* (2000), in their simulations, require that the elapsed time must exceed the cooling time, as calculated at a distance r_0 from the centre of the protogalaxy (*i.e.* at the edge of the core). This criterion avoids giving a false positive in the case in which $t_{\rm cool}$ drops briefly below $t_{\rm dyn}$ at an early time, and then climbs above it again as the H_2 abundance falls. I adopt a similar criterion in my model. Additionally, I have run a few simulations to examine the effect of applying a similar criterion at a much larger distance from the core – this is more appropriate if protogalactic cooling is primarily determined by the behaviour of the gas immediately behind the accretion shock, as suggested by the simulations of Fuller & Couchman (2000).

4.5 Self-consistent models of radiative feedback

Negative feedback

Before considering the main topic of this chapter – the effects of X-rays on protogalactic cooling – I make a slight digression to look at the effects of purely negative feedback and the uncertainty that is introduced by my approximate treatment of the H_2 rotational levels.

Comparing the results of simulations run using my two approximations – that the excited levels of H_2 are not populated, or that they have LTE populations – I find that at low flux levels, there is little or no difference to the results. For example, I plot in figure 4.3 the evolution with time of the H_2 abundance in the core of a protogalaxy with virial temperature $T_{\rm vir} = 10^3\,{\rm K}$ and collapse redshift $z = 30$, illuminated by a Lyman-Werner background with $J(\bar{\nu}) = 10^{-23}\,{\rm erg\,s^{-1}\,cm^{-2}\,Hz^{-1}\,sr^{-1}}$. [5] The solid line shows the behaviour if we ignore the excited levels; the dashed line the behaviour if LTE abundances are assumed. As we can see, there is no signif-

[5] Recall that, in the optically thin limit, $k_{\rm dis} = 1.38 \times 10^9\, J(\bar{\nu})\,{\rm s^{-1}}$.

icant difference between the two. This is not particularly surprising, since if the Lyman-Werner flux is small, then so is the photodissociation rate and substantial differences in it will have little effect on the final H_2 abundance.

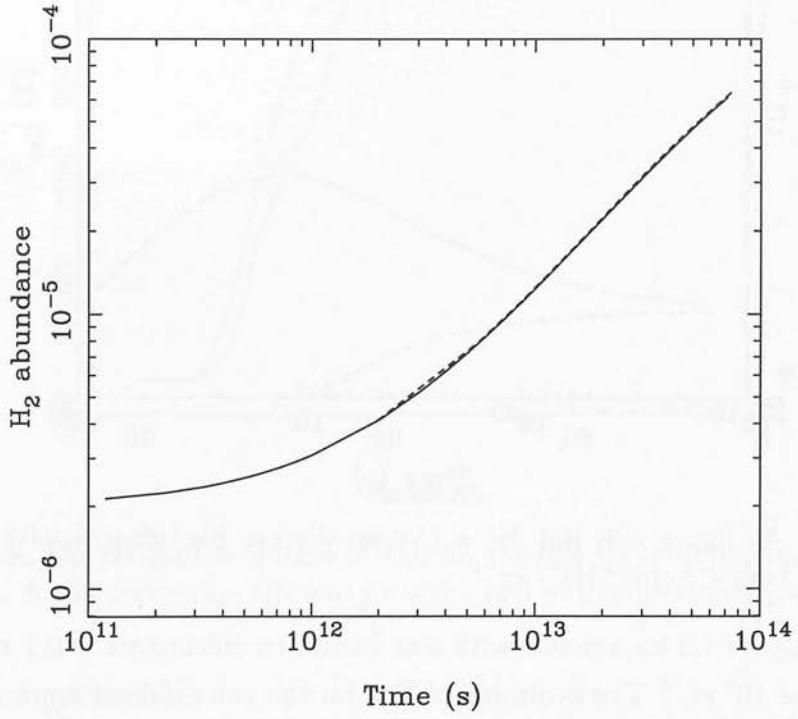

Figure 4.3: The evolution of the H_2 abundance in the core of a protogalaxy for two different H_2 level population approximations – no excited levels (solid line) or LTE populations (dashed line). The results plotted are for a protogalaxy with virial temperature $T_{\rm vir} = 10^3$ K and formation redshift $z = 30$, illuminated by a Lyman-Werner background with strength $J(\bar{\nu}) = 10^{-23}$ erg s^{-1} cm^{-2} Hz^{-1} sr^{-1}.

At higher flux levels, the difference in behaviour is more striking. For example, I plot in figure 4.4 the results of a similar comparison for a background with $J(\bar{\nu}) = 10^{-21}$. In this case, there is a clear qualitative difference in behaviour – in the ground state simulation, the H_2 abundance increases initially, before subsequently decreasing as the free electron fraction falls, while in the LTE simulation it falls right from the start. Nevertheless, despite the striking difference in the chemical evolution, the end result of the two simulations is the same – the protogalaxy fails to cool. This suggests that we may get a better idea of the importance of photodissociation from excited states if we compare the evolution of $T_{\rm crit}$ in the two models.

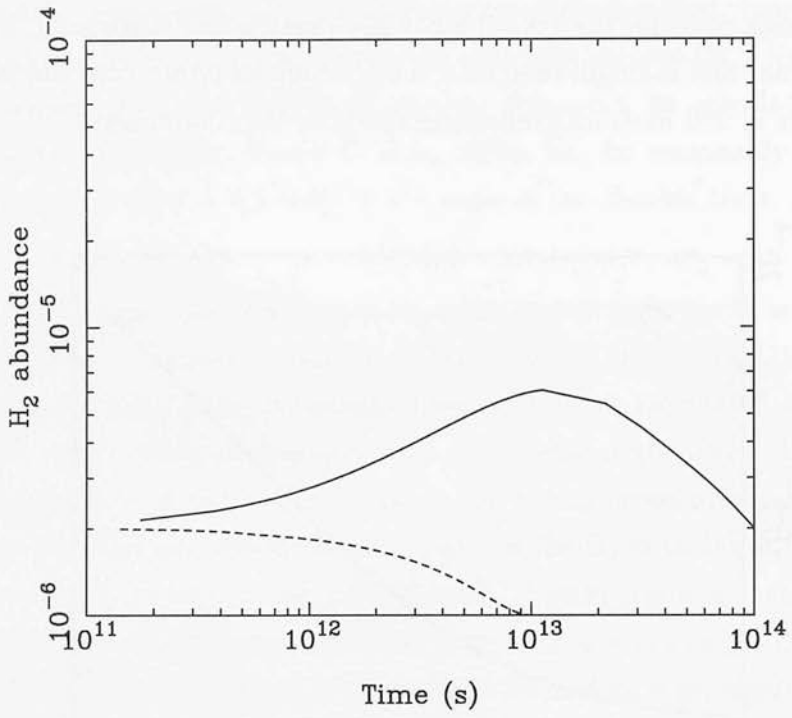

Figure 4.4: As figure 4.3, but for a Lyman-Werner background with strength $J(\bar{\nu}) = 10^{-21}$ erg s^{-1} cm^{-2} Hz^{-1} sr^{-1}.

I do this in figure 4.5 for a model with star formation efficiency $\epsilon = 0.1$ and source lifetime $t_{\rm on} = 10^7$ yr. [6] The evolution of $T_{\rm crit}$ for the two different approximations is broadly similar, with the onset of negative feedback occuring at slightly higher redshift in the LTE simulations.

An alternative way of looking at these results is to ask how strong the Lyman-Werner background must be for negative feedback to occur – unlike the redshift, this should be insensitive to details of the cosmological model. In figure 4.6 I plot the evolution of $T_{\rm crit}$ as a function of the mean Lyman-Werner flux in the two models. Again, we see that the evolution is broadly similar, with the critical flux differing by a factor of two at most. These results suggest that the degree to which we model the H_2 level populations will have some effect on the values of $T_{\rm crit}$ that we find, but that this effect is relatively small; as we shall see below, the effect of the X-ray background is far greater.

[6] Cf. figure 7 in Haiman *et al.* (2000), which shows the results of a similar model.

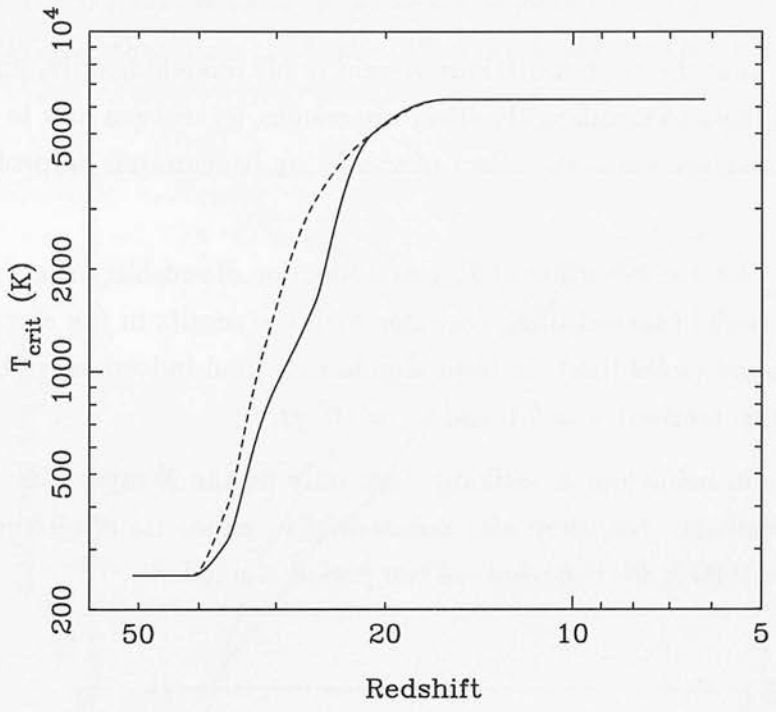

Figure 4.5: The redshift evolution of $T_{\rm crit}$ for the two H_2 level population approximations. A star formation efficiency $\epsilon = 0.1$ and source lifetime $t_{\rm on} = 10^7$ yr were assumed.

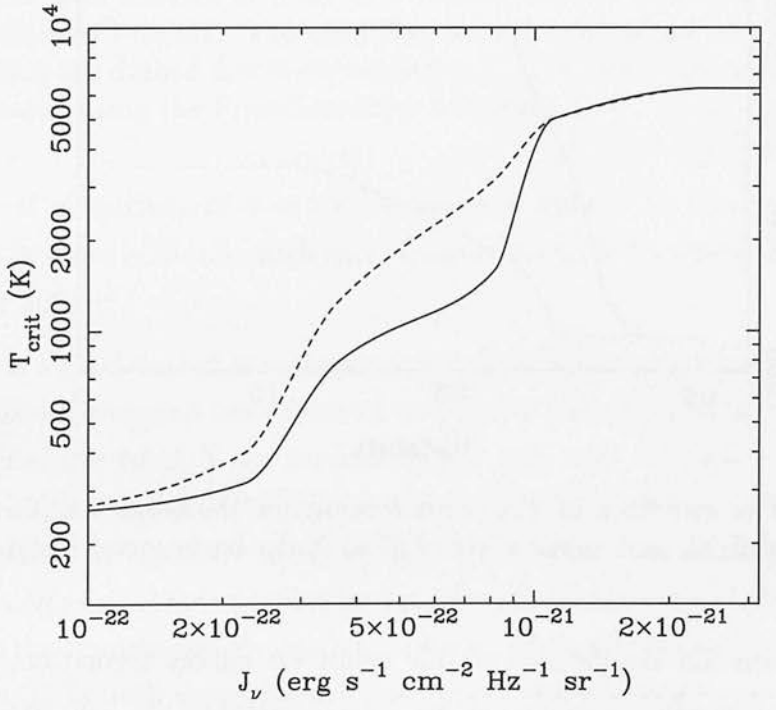

Figure 4.6: The same results plotted as a function of the mean Lyman-Werner background flux.

The empirical X-ray model

Having shown that the approximations present in my modelling of H_2 photodissociation are unlikely to significantly affect my results, let us turn now to the main topic of my investigation – the effect of the X-ray background on protogalactic cooling.

In figure 4.7 I plot the evolution of $T_{\rm crit}$ as a function of redshift for my empirical X-ray source model (dashed line), together with the results in the absence of an X-ray background (solid line). In both simulations (and indeed, all of the others presented in this section), $\epsilon = 0.1$ and $t_{\rm on} = 10^7$ yr.

The difference in behaviour is striking – not only do the X-rays delay the onset of negative feedback, but they also substantially reduce its effectiveness; $T_{\rm crit}$ remains below 1000 K for the whole of the period studied. [7]

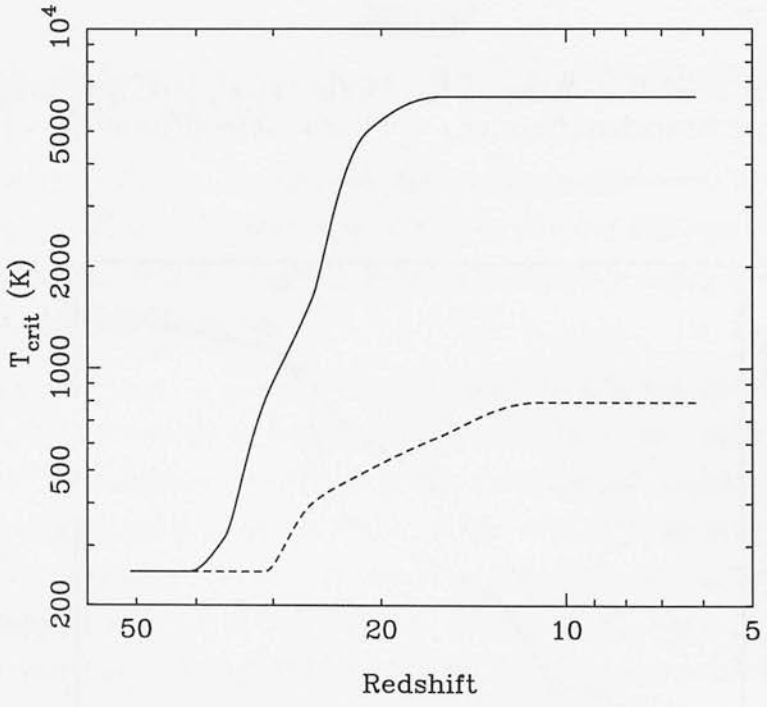

Figure 4.7: The evolution of $T_{\rm crit}$ with redshift for the empirical X-ray source model (dashed line), and in the absence of an X-ray background (solid line).

To demonstrate the significance of this result for galaxy formation, I plot in

[7] Although note that the effects of reionization – not included here – may alter this conclusion at low redshift.

figure 4.8 the fraction of mass in collapsed objects with $T_{\rm vir} > T_{\rm crit}$ for the two models. We see that for $z \lesssim 35$, the difference is substantial, being as much as

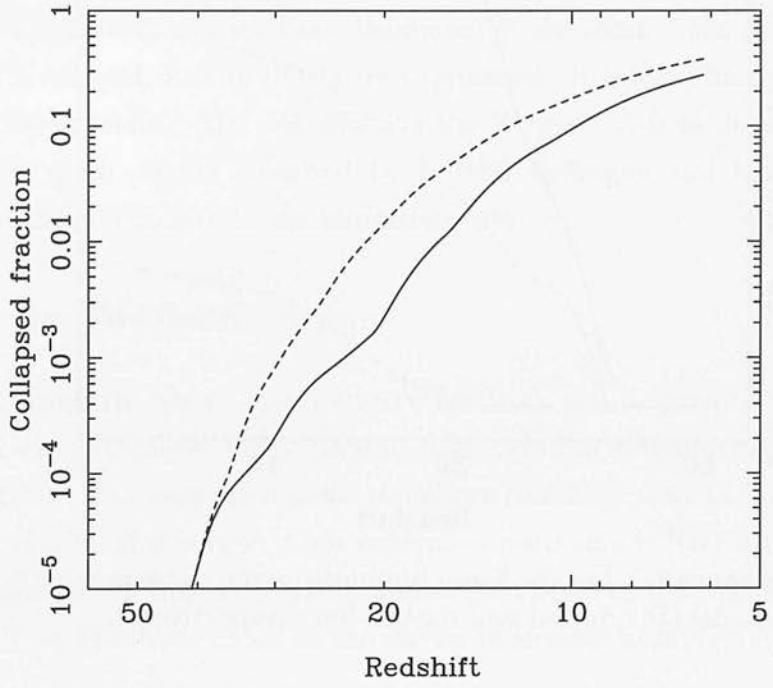

Figure 4.8: The fraction of mass in collapsed objects with $T_{\rm vir} > T_{\rm crit}$, plotted as a function of redshift. The solid line corresponds to the case with no X-ray background; the dashed line to my empirical X-ray model. The collapsed fractions are estimated using the Press-Schechter formalism.

an order of magnitude at $z \sim 20$. At lower redshifts, the two models begin to converge as more and more high mass galaxies form, but there is still a difference of almost 50% at $z = 5$.

As there is a good deal of uncertainty in the data underlying my empirical model, I have also investigated the effects of varying its basic parameters – the normalization (*i.e.* the total X-ray luminosity per unit star formation rate) and the spectral index. In figure 4.9 I show the results of increasing or decreasing the normalization by an order of magnitude. The effects are exactly as we would expect: increasing the normalization (and hence the strength of the background) increases the suppression of negative feedback, while decreasing the normalization decreases it. In both cases, however, the effect remains large

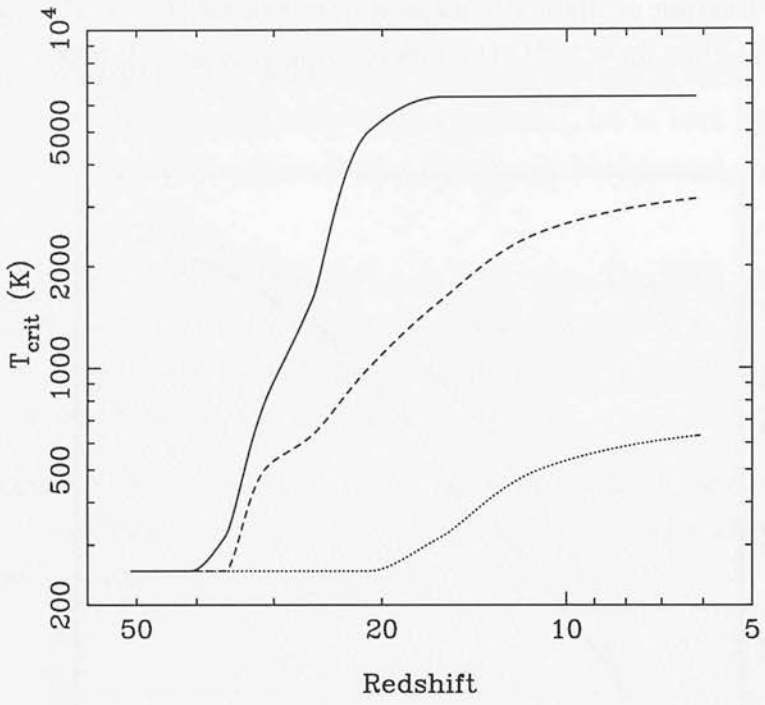

Figure 4.9: As figure 4.7, but for X-ray luminosities ten times greater or less than the standard model (the dotted and dashed lines respectively).

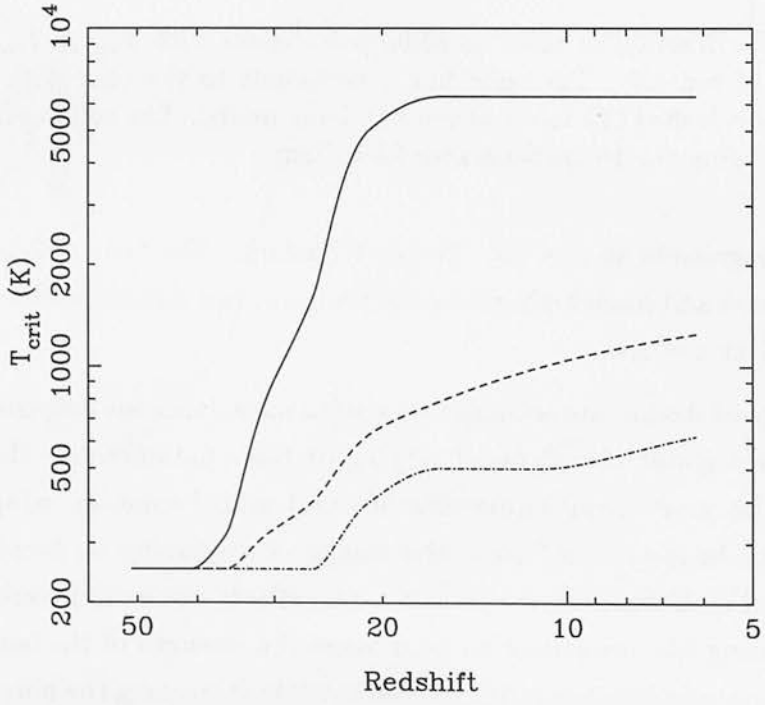

Figure 4.10: As figure 4.7, but for X-ray spectral indices $\Gamma = 1.2$ (dashed line) and $\Gamma = 1.8$.

In figure 4.10, I plot the results of simulations in which the normalization was kept constant and the spectral index was varied; results are plotted for $\Gamma = 1.2$ (dashed line) and $\Gamma = 1.8$ (dot-dashed line), corresponding to 1σ deviations from the mean value of Rephaeli *et al.* (1995). Some differences from the empirical model are apparent, but the basic behaviour is the same – the onset of negative feedback is delayed, and its effects are suppressed. It is also clear that the harder the emitted spectrum, the less effective the X-rays. This is unsurprising – soft X-rays are more readily absorbed by neutral hydrogen and helium, and thus contribute more strongly to the ionization rate.

Supernovae: Bremsstrahlung

It is clear from the above that negative feedback will be greatly reduced in any model in which massive X-ray binaries dominate the emission. However, as discussed previously, there are reasons to believe that they may be far less prevalent at high redshift. If this is so, then supernova remnants (SNR) will dominate the X-ray emission. In this section, I examine the effects of thermal bremsstrahlung from SNR; in the next, I look at the effects of inverse Compton emission.

In figure 4.11, I plot the evolution of $T_{\rm crit}$ for two thermal bremsstrahlung models. In one (given by the dot-dashed line), all of the emission comes from typical SNR, with characteristic temperatures $T_x \sim 1$ keV and which radiate a fraction $f_x \simeq 2 \times 10^{-4}$ of the their total energy as X-rays. The other model (the dashed line) assumes that all of the supernovae explode in very dense surroundings and form ultra-compact remnants, with $T_x \sim 30$ keV and $f_x \sim 0.01$. Any realistic model should lie between these limits.

We see from figure 4.11 that, compared to the empirical model, the effect of the X-rays is very much reduced – negative feedback *is* suppressed, but to a much smaller extent, and only for relatively large protogalaxies. This difference is borne out if we compare the collapsed mass fraction in the two models with that for the zero X-ray case (see figure 4.12). A small difference is apparent at $z \sim 20$, but has disappeared by $z = 5$.

It is also clear that there is little difference between the two supernova models, despite the large difference in f_x. This is due to the very hard spectra of the ultra-compact remnants – a substantial fraction of their radiation is at high photon energies ($h\nu > 1$ keV), little of which is absorbed within the protogalaxy.

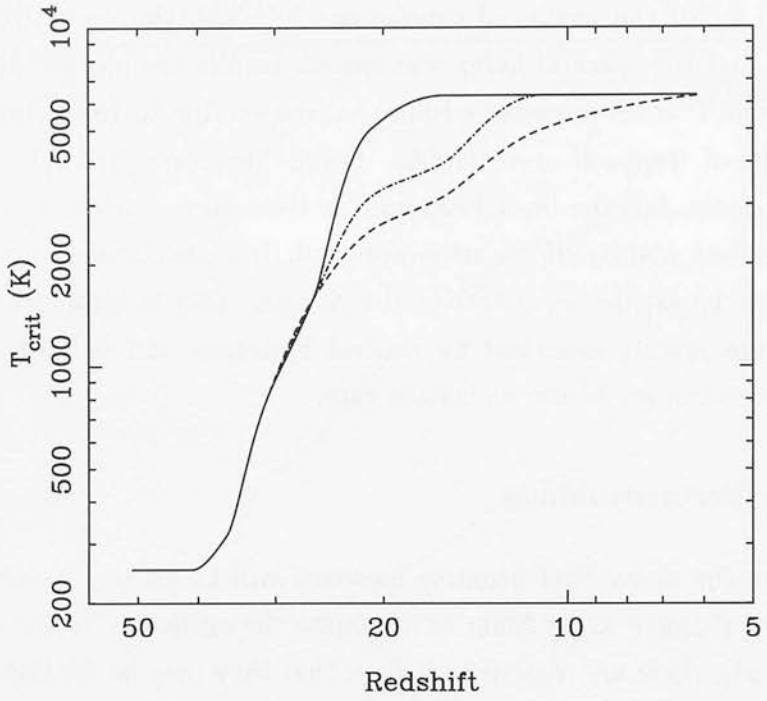

Figure 4.11: The evolution of $T_{\rm crit}$ for the thermal bremsstrahlung models. The dot-dashed line is for a model with 'typical' supernova remnants, the dashed for one with ultra-compact remnants.

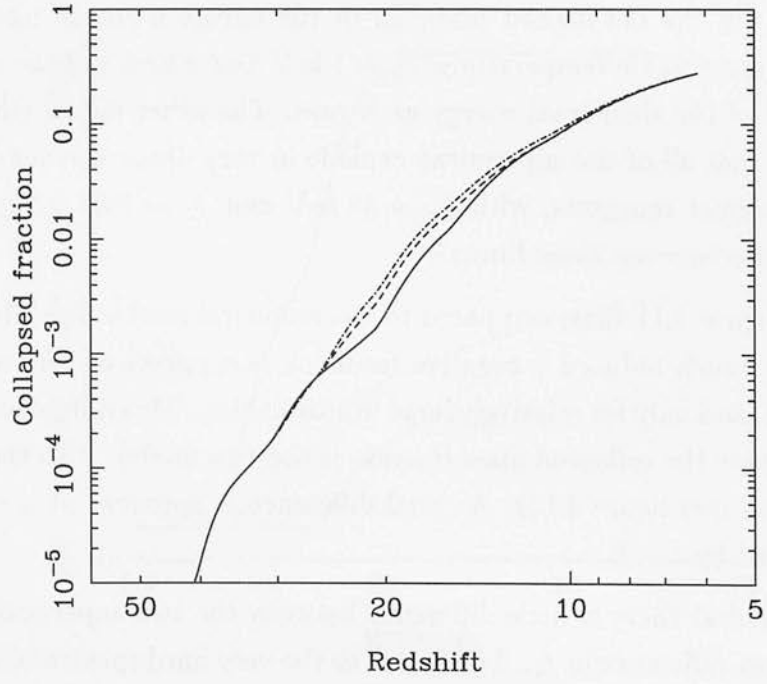

Figure 4.12: A similar comparison to that in figure 4.8, but for the two thermal bremsstrahlung models. The difference by $z = 5$ is negligible.

Supernovae: Inverse Compton emission

Plotted in figure 4.13 are the results for the two inverse Compton models: one in which 0.1% of the supernova energy is transferred into relativistic electrons (and thence to X-rays) and one in which as much as 10% is transferred; the true value likely lies somewhere between these bounds. We see immediately that these models are more effective at reducing negative feedback than the bremsstrahlung models, although they are still not as effective as the models based on X-ray binaries.

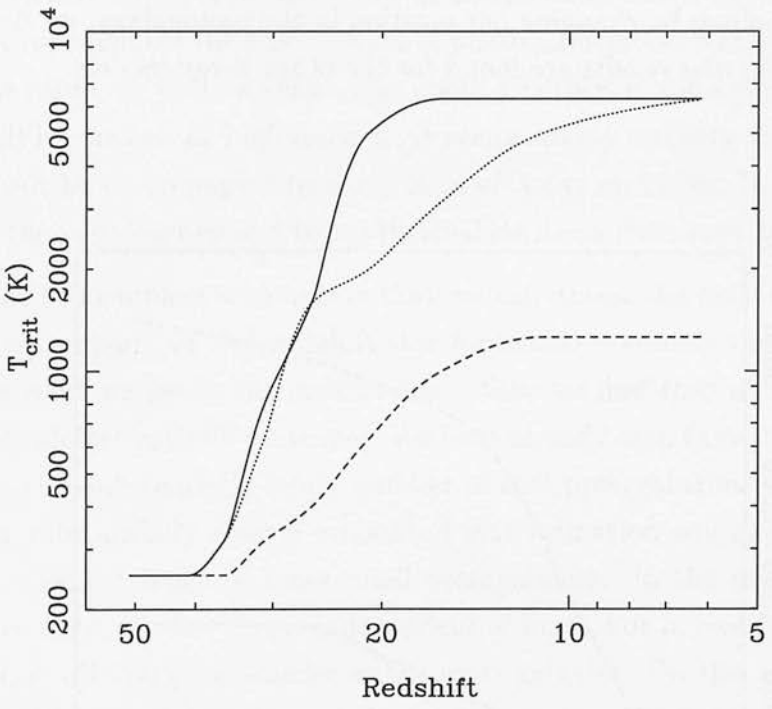

Figure 4.13: The redshift evolution of $T_{\rm crit}$ for inverse Compton scattering models in which 0.1% or 10% of the supernova energy is transferred to relativistic electrons (the dotted and dashed lines respectively).

Cooling outside the core

All of the results presented so far assume that whether or not a protogalaxy can cool is determined by the behaviour of the gas in its core regions. This is the assumption adopted by Haiman *et al.* (2000) and a good case can be made for it; nevertheless, it is still an assumption. A rather different assumption, motivated

by the results of Fuller & Couchman (2000), is that the fate of the protogalactic gas is determined substantially further out in the protogalaxy, in the regions lying just behind the accretion shock. To examine the effects of this alternative assumption, I have run simulations in which the cooling criterion is applied to gas at a distance $r = 7r_0$ from the centre of the protogalaxy; this is approximately one quarter of the distance to the truncation radius. In figure 4.14 I plot the results obtained for the empirical X-ray model using the two different cooling criteria.

It is immediately apparent that the degree to which the X-ray background acts to mitigate negative feedback is very dependent on the assumptions we make concerning the cooling; by changing our position in the protogalaxy, we significantly change $T_{\rm crit}$. Similar results are found for the other X-ray models.

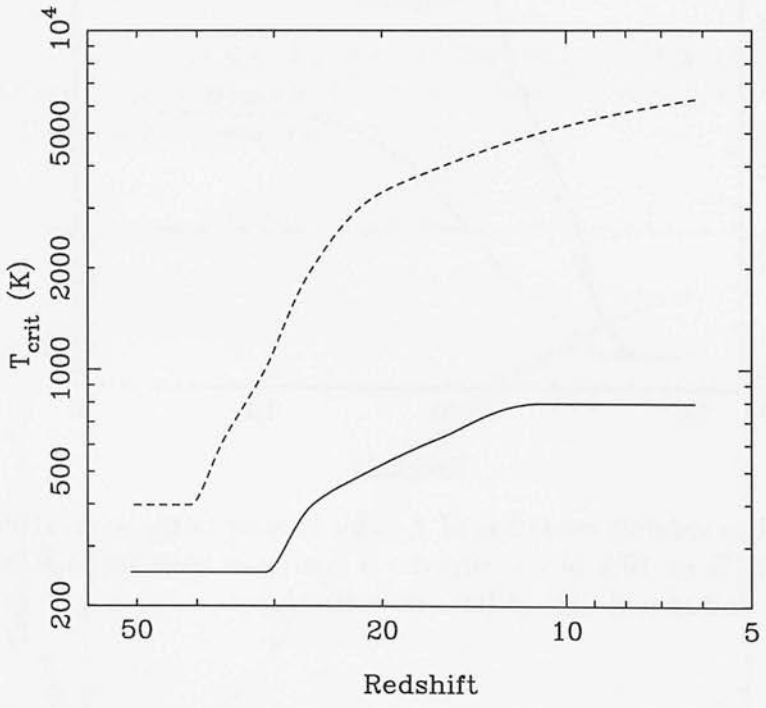

Figure 4.14: The effect of changing the cooling criterion. In both cases, the X-ray background is generated by my empirical source model. The solid line gives the evolution of $T_{\rm crit}$ when the effectiveness of cooling is determined at $r = r_0$; the dashed line corresponds to determination at $r = 7r_0$.

4.6 Conclusions

What conclusions can we draw from these results? To begin with, it should not be overlooked that I have successfully reproduced the results of Haiman *et al.* (2000) for the zero X-ray case, despite differences in both the chemical and computational models. This suggests that this approach is reasonably robust, and increases my confidence in the basic correctness of the remainder of my results.

The main fruits of this chapter, however, are obviously the results for the various X-ray source models. The precise effect of the X-ray background on the evolution of $T_{\rm crit}$ is clearly model dependent, but in *all* the cases that I have examined, including X-rays reduces the effectiveness of photodissociation feedback. This is an important result, as while we can argue about whether or not a particular type of source will be present at high redshift, it seems highly unlikely that star formation will not be accompanied by some level of X-ray emission. In particular, we would at the very least expect to see thermal emission from supernova remnants.

If we make the simplest assumption that we can about the nature of the X-ray emission accompanying high-redshift star formation – namely that it is broadly similar to what we see at the present day – then we find that photodissociation feedback is almost entirely overcome. We have already seen that this leads to the formation of a substantially larger number of cool protogalaxies; whether it also leads to a substantially greater amount of star formation will depend upon the star formation efficiency of these small protogalaxies. In the models presented here I have assumed that this is independent of mass, but in reality, as we saw in chapter 3, it will likely be smaller in low-mass galaxies. On the other hand, the ease with which ionizing photons and metals can escape from small protogalaxies suggests that star formation within them will have a disproportionately large influence on the IGM.

These conclusions all presume, of course, that the cooling of gas within the protogalaxy is ultimately controlled by the behaviour of the gas in the core regions. This is a reasonable assumption, but an assumption nevertheless, and it is worrying that making a different assumption – that efficient cooling is determined by the behaviour of the gas in the halo region – leads to very different results. This immediately suggests an obvious follow-up to the work presented in this chapter – combine the chemical and radiative transfer models presented here with a proper hydrodynamical simulation. This would avoid our cooling criterion

problem, as we could simply measure the amount of gas which cools, and would also remove the need for many of our other approximations. It does, however, involve overcoming a large number of practical problems, and is clearly work for the future.

Bibliography

Abel, T., Anninos, P., Zhang, Y., Norman, M. L., 1997, New Astron., **2**, 181.

Abel, T., Bryan, G. L., Norman, M. L., 2000, ApJ, **540**, 39.

Abel, T., Bryan, G. L., Norman, M. L., 2001, In Preparation.

Abgrall, H., Bourlot, J. L., des Forêts, G. P., Roueff, E., Flower, D. R., Heck, L., 1992, A&A, **253**, 525.

Abgrall, H., Roueff, E., Launay, F., Roncin, J., Subtil, J., 1993, A&AS, **101**, 273.

Abgrall, H., Roueff, E., Launay, F., Roncin, J., Subtil, J., 1993, A&AS, **101**, 323.

Abgrall, H., Roueff, E., Drira, I., 2000, A&AS, **141**, 297.

Adams, T. F., 1976, A&A, **50**, 461.

Adler, R. J., 1981. *The Geometry of Random Fields.* Wiley.

Aggarwal, K. M., Callaway, J., Kingston, A. E., Unnikrishnan, K., 1992, ApJS, **80**, 473.

Aldrovandi, S. M. V., Pequignot, D., 1973, A&A, **25**, 137.

Allison, A. C., Dalgarno, A., 1969, At. Data, **5**, 91.

Anninos, P., Zhang, Y., Abel, T., Norman, M. L., 1997, New Astron., **2**, 209.

Aretxaga, I., Benetti, S., Terlevich, R. J., Fabian, A. C., Cappellaro, E., Turatto, M., della Valle, M., 1999, MNRAS, **309**, 343.

Auer, L., Heasley, J., 1971, Yale University Observatory Report, unpublished.

Bagla, J. S., Padmanabhan, T., 1994, MNRAS, **266**, 227.

Baraffe, I., Heger, A., Woosley, S. E., 2001, ApJ, **550**, 890.

Bardeen, J. M., Bond, J. R., Kaiser, N., Szalay, A. S., 1986, ApJ, **304**, 15.

Barkana, R., Loeb, A., 1999, ApJ, **523**, 54.

Bates, D. R., Öpik, U., 1968, J. Phys. B, **1**, 543.

Beers, T. C., in *The First Stars*, eds. A. Weiss, T. Abel, and V. Hill, Springer, 2000.

Bertoldi, F., 1989, ApJ, **346**, 735.

Bertoldi, F., McKee, C. F., 1990, ApJ, **354**, 529.

Bertschinger, E., 1998, ARA&A, **36**, 599.

Birkhoff, G. D., 1923. *Relativity and Modern Physics.* Harvard University Press.

Black, J. H., 1978, ApJ, **222**, 125.

Black, J. H., 1981, MNRAS, **197**, 553.

Black, J. H., in *Molecular Astrophysics*, ed. T. W. Hartquist, page 473, CUP, 1991.

Black, J. H., Dalgarno, A., 1977, ApJS, **34**, 405.

Blumenthal, G. R., Faber, S. M., Primack, J. R., Rees, M. J., 1984, Nature, **311**, 517.

Bond, J. R., Efstathiou, G., 1984, ApJ, **285**, L45.

Bond, J. R., Efstathiou, G., Silk, J., 1980, Phys. Rev. Lett., **45**, 1980.

Bond, J. R., Cole, S., Efstathiou, G., Kaiser, N., 1991, ApJ, **379**, 440.

Boothroyd, A. I., Keogh, W. J., Martin, P. G., Peterson, M. R., 1996, J. Chem. Phys., **104**, 7139.

Bouchet, F. R., in *Dark Matter in the Universe*, IOS Press, 1996.

Bray, I., Burgess, A., Fursa, D. V., Tully, J. A., 2000, A&AS, **146**, 481.

Brocklehurst, M., 1971, MNRAS, **153**, 471.

Bromm, V., Coppi, P. S., Larson, R. B., 2002, ApJ, **564**, 23.

Browne, J. C., Dalgarno, A., 1969, J. Phys. B., **2**, 885.

Bruzual, G. A., Charlot, S., 1993, ApJ, **405**, 538.

Bryan, G. L., Norman, M. L., 1998, ApJ, **495**, 80.

Burles, S., Nollett, K. M., Turner, M. S., 2000, astro-ph/0008495.

Burles, S., Nollett, K. M., Turner, M. S., 2000, ApJ, **552**, L1.

Caldwell, R. R., Dave, R., Steinhardt, P. J., 1998, Ap & SS, **261**, 303.

Callaway, J., 1994, ADNDT, **57**, 10.

Carroll, S. M., Press, W. H., Turner, E. L., 1992, ARA&A, **30**, 499.

Cassisi, S., Castellani, V., 1993, ApJS, **88**, 509.

Cen, R., 1992, ApJS, **78**, 341.

Chandrasekhar, S., 1943, Rev. Mod. Phys., **15**, 2.

Chevalier, R. A., 1999, ApJ, **511**, 798.

Ciardi, B., Ferrara, A., Governato, F., Jenkins, A., 2000, MNRAS, **314**, 611.

Ciardi, B., Ferrara, A., Abel, T., 2000, ApJ, **533**, 594.

Cojazzi, P., Bressan, A., Lucchin, F., Pantano, O., Chavez, M., 2000, MNRAS, **315**, L51.

Colberg, J. M. *et al.*, 2000, MNRAS, **319**, 209.

Cole, S., Lacey, C., 1996, MNRAS, **281**, 716.

Cowie, L. L., Songaila, A., 1998, Nature, **394**, 44.

Cox, D. P., 1983, ApJ, **265**, L61.

Culhane, M., McCray, R., 1995, ApJ, **455**, 335.

Dahlem, M., Parmar, A., Oosterbroek, T., Orr, A., Weaver, K. A., Heckman, T. M., 2000, ApJ, **538**, 555.

Dalgarno, A., Lepp, S., in *Astrochemistry*, eds. M. S. Vardya and S. P. Tarafdar, Reidel, 1987.

Dalgarno, A., McDowell, M. R. C., 1956, Proc. Phys. Soc. London A, **69**, 615.

Dalgarno, A., Yan, M., Liu, W., 1999, ApJS, **125**, 237.

Datz, S., Sundström, G., Biedermann, C., Brostöm, L., Danared, H., Mannervik, S., Rowat, J. R., Larsson, M., 1995, Phys. Rev. Lett., **74**, 896.

David, L. P., Jones, C., Forman, W., 1992, ApJ, **388**, 82.

De Bernadis, P. *et al.*, 2000, Nature, **404**, 955.

de Jong, T., 1972, A&A, **20**, 263.

di Nella, H., Montuori, M., Paturel, G., Pietronero, L., Sylos Labini, F., 1996, A&A, **308**, L33.

Draine, B. T., Bertoldi, F., 1996, ApJ, **468**, 269.

Drake, G. W. F., in *Atomic, Molecular and Optical Physics Handbook*, ed. G. W. F. Drake, American Institute of Physics, Woodbury, New York, 1996.

Dunn, G. H., 1968, Phys. Rev., **172**, 1.

Efstathiou, G., in *Physics of the Early Universe*, eds. J. A. Peacock, A. F. Heavens, and A. T. Davies, page 361, SUSSP, 1990.

Efstathiou, G., 1992, MNRAS, **256**, 43P.

Ellison, S. L., Songaila, A., Schaye, J., Pettini, M., 2000, AJ, **120**, 1167.

Elmegreen, B. G., in *Origins of Galaxies, Stars, Planets and Life*, eds. C. E. Woodward, H. A. Thronson, and M. Shull, page 150, ASP, 1998.

Emerson, D., 1996. *Interpreting Astronomical Spectra.* Wiley.

Fan, X. *et al.*, 2000, AJ, **120**, 1167.

Ferland, G. J., Peterson, B. M., Horne, K., Welsh, W. F., Nahar, S. N., 1992, ApJ, **387**, 95.

Ferrara, A., Tolstoy, E., 2000, MNRAS, **313**, 291.

Field, G. B., Somerville, W. B., Dressler, K., 1966, ARA&A, **4**, 207.

Fischer, D. A., Marcy, G. W., 1992, ApJ, **396**, 178.

Fixsen, D. J., Cheng, E. S., Gales, J. M., Mather, J. C., Shafer, R. A., Wright, E. L., 1996, ApJ, **473**, 576.

Flower, D. R., 1997, J. Phys. B, **30**, 3009.

Flower, D. R., Pineau des Forêts, G., 2000, MNRAS, **316**, 901.

Flower, D. R., Roueff, E., 1998, J. Phys. B, **31**, L955.

Flower, D. R., Roueff, E., 1998, J. Phys. B, **31**, 2935.

Flower, D. R., Roueff, E., Zeippen, C. J., 1998, J. Phys. B, **31**, 1105.

Forieri, C., 1982. Ph.D. thesis, Institute of Astronomy, University of Padova.

Franco, J., Tenorio-Tagle, G., Bodenheimer, P., 1990, ApJ, **349**, 126.

Friedmann, A., 1922, Z. Phys., **10**, 377.

Friedmann, A., 1924, Z. Phys., **21**, 326.

Fuller, T. M., Couchman, H. M. P., 2000, ApJ, **544**, 6.

Galli, D., Palla, F., 1998, A&A, **335**, 403.

Garcia-Segura, G., Franco, J., 1996, Rev. Mex. Astron. Conf., **40**, 98.

Gerlich, D., 1990, J. Chem. Phys., **92**, 2377.

Glass-Maujean, M., 1986, Phys. Rev. A, **33**, 342.

Glover, S. C. O., Brand, P. W. J. L., 2001, MNRAS, **321**, 385.

Gnedin, N. Y., Hui, L., 1998, MNRAS, **296**, 44.

Goldman, S. P., 1989, Phys. Rev. A, **40**, 1185.

Gould, R. J., Thakur, R. K., 1970, Ann. Phys., **61**, 351.

Griffiths, L. M., Barbosa, D., Liddle, A. R., 1999, MNRAS, **308**, 854.

Guberman, S. L., 1994, Phys. Rev. A, **49**, R4277.

Gurbatov, S. N., Saichev, A. I., Shandarin, S. F., 1989, MNRAS, **236**, 385.

Guth, A., Weinberg, E., 1983, Nucl. Phys. B, **212**, 321.

Guth, A. H., 1981, Phys. Rev. D, **23**, 347.

Haiman, Z., Thoul, A. A., Loeb, A., 1996, ApJ, **464**, 523.

Haiman, Z., Rees, M. J., Loeb, A., 1996, ApJ, **467**, 522.

Haiman, Z., Rees, M. J., Loeb, A., 1997, ApJ, **476**, 458.

Haiman, Z., Madau, P., Loeb, A., 1999, ApJ, **514**, 535.

Haiman, Z., Abel, T., Rees, M. J., 2000, ApJ, **534**, 11.

Haiman, Z., Abel, T., Madau, P., 2001, ApJ, **551**, 599.

Hanany, S. *et al.*, 2000, ApJ, **545**, L5.

Harrison, E. R., 1970, Phys. Rev. D, **1**, 2726.

Hatton, S., 1999, MNRAS, **310**, 1128.

Heiles, C., 1971, ARA&A, **9**, 293.

Helfand, D. J., Moran, E. C., 2001, ApJ, **554**, 27.

Henry, J. P., Arnaud, K. A., 1991, ApJ, **372**, 410.

Hollenbach, D., Mckee, C. F., 1979, ApJS, **41**.

Holliday, M. G., Muckerman, J. T., Friedman, L., 1971, J. Chem. Phys., **54**, 1058.

Hu, W., Sugiyama, N., 1995, ApJ, **444**, 489.

Hu, W., Scott, D., Sugiyama, N., White, M., 1995, Phys. Rev. D, **52**, 5498.

Hubble, E., 1929, Proc. N. A. S., **15**, 168.

Hubeny, I., Lanz, T., 1995, ApJ, **439**, 875.

Hubeny, I., Lanz, T., Heap, S., in *The First Stars*, eds. A. Weiss, T. Abel, and V. Hill, Springer, 2000.

Hui, L., Gnedin, N. Y., 1997, MNRAS, **292**, 27.

Hummer, D. G., Storey, P. J., 1998, MNRAS, **297**, 1073.

Huq, M. S., Doverspike, L. D., Champion, R. L., Esaulov, V. A., 1982, J. Phys. B., **15**, 951.

Huq, M. S., Doverspike, L. D., Champion, R. L., 1983, Phys. Rev. A., **27**, 2831.

Janev, R. K., Langer, W. D., Evans, K., Post, D. E., 1987. *Elementary Processes in Hydrogen-Helium Plasmas.* Springer.

Jang-Condell, H., Hernquist, L., 2001, ApJ, **548**, 68.

Jeans, J., 1902, Phil. Trans. R. Soc. A, **199**, 67.

Juřek, M., Špirko, V., Kraemer, W. P., 1995, Chem. Phys., **193**, 287.

Kaellander, D., Hultman, J., 1998, A&A, **333**, 399.

Karpas, Z., Anicich, V., Huntress, W. T., 1979, J. Chem. Phys., **70**, 2877.

Kennicutt, R. C., 1998, ARA&A, **36**, 189.

Kepner, J. V., Babul, A., Spergel, D. N., 1997, ApJ, **487**, 61.

Kimura, M., Lane, N. F., Dalgarno, A., Dixson, R. G., 1993, ApJ, **405**, 801.

Kitayama, T., Ikeuchi, S., 2000, ApJ, **529**, 615.

Kolb, E. W., Turner, M. S., 1990. *The Early Universe.* Addison Wesley.

Koyama, K., Petre, R., Gotthelf, E. V., Hwang, U., Matsura, M., Ozaki, M., Holt, S. S., 1995, Nature, **378**, 255.

Kraemer, W. P., Špirko, V., Juřek, M., 1995, Chem. Phys. Lett., **236**, 177.

Kudritzki, R. P., in *The First Stars*, eds. A. Weiss, T. Abel, and V. Hill, Springer, 2000.

Lacey, C., Cole, S., 1993, MNRAS, **262**, 627.

Lacey, C., Cole, S., 1994, MNRAS, **271**, 676.

Lahav, O., Lilje, P. B., Primack, J. R., Rees, M. J., 1991, MNRAS, **251**, 128.

Launay, J. M., Le Dourneuf, M., Zeippen, C. J., 1991, A&A, **252**, 842.

Le Bourlot, J., Pineau des Forêts, G., Flower, D. R., 1999, MNRAS, **305**, 802.

Leitherer, C., Heckman, T. M., 1995, ApJS, **96**, 9.

Lenzuni, P., Chernoff, D. F., Salpeter, E. E., 1991, ApJS, **76**, 759.

Lepp, S., Shull, J. M., 1983, ApJ, **270**, 578.

Lesgourgues, J., Polarski, D., Prunet, S., Starobinsky, A. A., 2000, A&A, **359**, 414.

Liddle, A. R., in *From Quantum Fluctuations to Cosmological Structures*, eds. D. Valls-Gabaud, M. A. Hendry, P. Molaro, and K. Chamcham, page 31, ASP, 1997.

Liddle, A. R., Lyth, D. H., 1993, Phys. Reports, **231**, 1.

Linde, A., in *Cosmological Constant and the Evolution of the Universe*, eds. K. Sato, T. Suginohara, and T. Sugiyama, page 133, Universal Academy Press, 1996.

Linder, F., Janev, R. K., Botero, J., in *Atomic and Molecular Processes in Fusion Edge Plasmas*, ed. R. K. Janev, Plenum Press, 1995.

Loeb, A., Barkana, R., 2001, ARA&A, **39**, 19.

Mac Low, M., Ferrara, A., 1999, ApJ, **513**, 142.

Machacek, M. E., Bryan, G. L., Abel, T., 2001, ApJ, **548**, 509.

Madau, P., Haardt, F., Rees, M. J., 1999, ApJ, **514**, 648.

Maddox, S., in *Clustering at High Redshift*, eds. A. Mazure, O. Le Fèvre, and V. Le Brun, page 63, ASP, 2000.

Marigo, P., Girardi, L., Chiosi, C., Wood, P. R., 2001, A&A, **371**, 152.

Martin, P. G., Schwarz, D. H., Mandy, M. E., 1996, ApJ, **461**, 265.

Matarrese, S., Lucchin, F., Moscardini, L., Saez, D., 1992, MNRAS, **259**, 437.

Meier, D. L., 1976, ApJ, **207**, 343.

Meiksin, A., White, M., Peacock, J. A., 1999, MNRAS, **304**, 851.

Monaco, P., 1997, MNRAS, **287**, 753.

Monaco, P., 1998, Fund. Cosm. Phys., **19**, 157.

Moore, B., 2001, astro-ph/0103100.

Moore, B., Quinn, T., Governato, F., Stadel, J., Lake, G., 1999, MNRAS, **310**, 1147.

Moran, E. C., Lehnert, M. D., 1997, ApJ, **478**, 172.

Moran, E. C., Lehnert, M. D., Helfand, D. J., 1999, ApJ, **526**, 649.

Moseley, J., Aberth, W., Peterson, J. A., 1970, Phys. Rev. Lett., **24**, 435.

Navarro, J. F., Frenk, C. S., White, S. D. M., 1997, ApJ, **490**, 493.

Nishi, R., Susa, H., 1999, ApJ, **523**, L103.

Oh, S. P., 2001, ApJ, **553**, 499.

Olive, K. A., Steigman, G., Walker, T. P., 2000, Phys. Rep., **333-334**, 389.

O'Meara, J. M., Tytler, D., Kirkman, D., Suzuki, N., Prochaska, J. X., Lubin, D., Wolfe, A. M., 2001, ApJ, **552**, 718.

Omukai, K., Nishi, R., 1998, ApJ, **508**, 141.

Omukai, K., Nishi, R., 1999, ApJ, **518**, 64.

O'Neil, S. V., Reinhardt, W. P., 1978, J. Chem. Phys., **69**, 2126.

Osterbrock, D. E., 1989. *Astrophysics of Gaseous Nebulae and Active Galactic Nuclei.* University Science Books.

Ostriker, J. P., Gnedin, N. Y., 1996, ApJ, **472**, L63.

Ostriker, J. P., Steinhardt, P. J., 1995, Nature, **377**, 600.

Padmanabhan, T., 1993. *Structure Formation in the Universe.* CUP.

Pagel, B. E. G., 2000, Phys. Rep., **333-334**, 433.

Palla, F., Zinnecker, H., in *Starbursts and Galaxy Evolution*, eds. T. X. Thuan, T. Montmerle, and J. Tran Thanh Van, page 533, Editions Frontieres, 1987.

Palla, F., Salpeter, E. E., Stahler, S. W., 1983, ApJ, **271**, 632.

Peacock, J. A., 1999. *Cosmological Physics.* C.U.P.

Peacock, J. A., Heavens, A. F., 1990, MNRAS, **243**, 133.

Pearce, F. R., Couchman, H. M. P., 1997, New Astron., **2**, 411.

Peebles, P. J. E., 1974, A&A, **32**, 391.

Peebles, P. J. E., 1980. *The Large Scale Structure of the Universe.* Princeton.

Peebles, P. J. E., 1993. *Principles of Physical Cosmology.* Princeton University Press.

Pei, Y. C., 1995, ApJ, **438**, 623.

Perlmutter, S. *et al.*, 1999, ApJ, **517**, 565.

Peterson, J. A., Aberth, W. H., Moseley, J. T., Sheridan, J. R., 1971, Phys. Rev. A, **3**, 1651.

Petre, R., Allen, G. E., Hwang, U., 1999, Astr. Nach., **320**, 199.

Poulaert, G., Brouillard, F., Claeys, W., McGowan, J. W., Wassenhove, G. V., 1978, J. Phys. B, **11**, L671.

Press, W. H., Schechter, P., 1974, ApJ, **187**, 425.

Press, W. H., Flannery, B. P., Teukolsky, S. A., Vetterling, W. T., 1986. *Numerical Recipes: The Art of Scientific Computing.* CUP.

Ramaker, D. E., Peek, J. M., 1976, Phys. Rev. A, **13**, 58.

Rees, M. J., Ostriker, J. P., 1977, MNRAS, **179**, 541.

Rephaeli, Y., Gruber, D., Persic, M., 1995, A&A, **300**, 91.

Ricotti, M., Shull, J. M., 2000, ApJ, **542**, 548.

Riess, A. G. *et al.*, 1998, AJ, **116**, 1009.

Robertson, H. P., 1935, Proc. N. A. S., **15**, 822.

Rodgers, C. D., Williams, A. P., 1974, JQSRT, **14**, 319.

Roueff, E., 1997. private communication.

Salati, P., in *From Quantum Fluctuations to Cosmological Structures*, eds. D. Valls-Gabaud, M. A. Hendry, P. Molaro, and K. Chamcham, page 63, ASP, 1997.

Salpeter, E. E., 1955, ApJ, **121**, 161.

Samson, J. A. R., Haddad, G. N., 1994, J. Opt. Soc. Am. B, **11**, 277.

Sanders, D. B., Mirabel, I. F., 1996, ARA&A, **34**, 749.

Sarpal, B. K., Tennyson, J., 1993, MNRAS, **263**, 909.

Scannapieco, E., Ferrara, A., Broadhurst, T., 2000, ApJ, **536**, L11.

Schaeffer, R., Silk, J., 1988, ApJ, **332**, 1.

Schneider, I. F., Dulieu, O., Giusti-Suzor, A., Roueff, E., 1994, ApJ, **424**, 983.

Seager, S., Sasselov, D. D., Scott, D., 1999, ApJ, **523**, L1.

Seager, S., Sasselov, D. D., Scott, D., 2000, ApJS, **128**, 407.

Sedov, L. I., 1959. *Similarity and Dimensional Methods in Mechanics.* Academic Press.

Seljak, U., Zaldarriaga, M., 1996, ApJ, **469**, 437.

Shafi, Q., Stecker, F. W., 1984, Phys. Rev. Lett., **53**, 1292.

Shapiro, P. R., Kang, H., 1987, ApJ, **318**, 32.

Shapiro, P. R., Iliev, I. T., Raga, A. C., 1999, MNRAS, **307**, 203.

Sheth, R. K., Mo, H. J., Tormen, G., 2001, MNRAS, **323**, 1.

Sidhu, K. S., Miller, S., Tennyson, J., 1992, A&A, **255**, 453.

Silk, J., 1968, ApJ, **151**, 459.

Silk, J., Wilson, M. L., 1981, ApJ, **244**, L37.

Snow, T. P., 1973, ApJ, **184**, 135.

Spitzer, L., 1978. *Physical Processes in the Interstellar Medium.* Wiley.

Stancil, P. C., 1994, ApJ, **430**, 360.

Stancil, P. C., Babb, J. F., Dalgarno, A., 1993, ApJ, **414**, 672.

Stancil, P. C., Lepp, S., Dalgarno, A., 1998, ApJ, **509**, 1.

Stecher, T. P., Williams, D. A., 1967, ApJ, **149**, L29.

Stibbe, D. T., Tennyson, J., 1999, ApJ, **513**, L147.

Strickland, D. K., Stevens, I. R., 2000, MNRAS, **314**, 511.

Sugiyama, N., 1995, ApJS, **100**, 281.

Tegmark, M., 1994. Ph.D. thesis, Univ. California Berkeley.

Tegmark, M., Silk, J., Rees, M., Blanchard, A., Abel, T., Palla, F., 1997, ApJ, **474**, 1.

Terlevich, R., Tenorio-Tagle, G., Franco, J., Melnick, J., 1992, MNRAS, **255**, 713.

Thacker, R. J., Couchman, H. M. P., 2000, ApJ, **545**, 728.

Thoul, A. A., Weinberg, D. H., 1996, ApJ, **465**, 608.

Tiné, S., Lepp, S., Dalgarno, A., 1998, Mem. Soc. Astron. It., **69**, 345.

Trentham, N., Möller, O., Ramirez-Ruiz, E., 2001, MNRAS, **322**, 658.

Truelove, J. K., Klein, R. I., McKee, C. F., Holliman, J. H., Howell, L. H., Greenough, J. A., Woods, D. T., 1998, ApJ, **495**, 821.

Tumlinson, J., Shull, J. M., 2000, ApJ, **528**, L65.

Tytler, D., O'Meara, J. M., Suzuki, N., Lubin, D., 2000, Physica Scripta, **85**, 12.

Valentine, H., Saunders, W., Taylor, A., 2000, MNRAS, **319**, L13.

van Kampen, E., Jimenez, R., Peacock, J. A., 1999, MNRAS, **310**, 43.

van Paradijs, J., McClintock, J., in *X-ray Binaries*, eds. Lewin, van Paradijs, and van den Heuvel, page 58, CUP, 1995.

Voronov, G. S., 1997, ADNDT, **65**, 1.

Walker, A. G., 1936, Proc. London Math. Soc., **42**, 90.

Wang, L., Caldwell, R. R., Ostriker, J. P., Steinhardt, P. J., 2000, ApJ, **530**, 17.

Weinberg, S., 1989, Rev. Mod. Phys., **61**, 1.

Weiss, A., Cassisi, S., Schlattl, H., Salaris, M., 2000, ApJ, **533**, 413.

Wilms, J., Allen, A., McCray, R., 2000, ApJ, **542**, 914.

Wirtz, C., 1924, Astron. Nachr., **222**, 21.

Wishart, A. W., 1979, MNRAS, **187**, 59P.

Wolniewicz, L., Simbotin, I., Dalgarno, A., 1998, ApJS, **115**, 293.

Worthey, G., 1994, ApJS, **95**, 107.

Wu, K. K. S., Lahav, O., Rees, M. J., 1999, Nature, **397**, 225.

Yan, M., Sadeghpour, H. R., Dalgarno, A., 1998, ApJ, **496**, 1044.

York, D. G. *et al.*, 2000, AJ, **120**, 1579.

Yorke, H. W., 1986, ARA&A, **24**, 49.

Zeldovich, Y. B., 1967, J.E.T.P. Lett., **6**, 316.

Zeldovich, Y. B., 1970, A&A, **5**, 84.

Zeldovich, Y. B., 1972, MNRAS, **160**, 1P.

Zezas, A. L., Georgantopolous, I., Ward, M. J., 1998, MNRAS, **301**, 915.

Zygelman, B., Dalgarno, A., Kimura, M., Lane, N. F., 1989, Phys. Rev. A, **40**, 2340.